DOCUMENT RESUME

ED 431 632 SE 062 700

AUTHOR Klein, Stephen; Hamilton, Laura; McCaffrey, Daniel; Stecher, Brian; Robyn, Abby; Burroughs, Delia

TITLE Teaching Practices and Student Achievement: Report of First-Year Findings from the "Mosaic" Study of Systemic Initiatives in Mathematics and Science.

INSTITUTION Rand Corp., Santa Monica, CA.
SPONS AGENCY National Science Foundation, Arlington, VA.
REPORT NO DRU-2083-EDU
PUB DATE 1999-05-00
NOTE 100p.
PUB TYPE Reports - Research (143)
EDRS PRICE EDRS Price MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS *Academic Achievement; Educational Assessment; Educational Change; Educational Innovation; Elementary Secondary Education; *Mathematics Achievement; *Mathematics Instruction; Program Effectiveness; *Science Instruction; Teacher Education
IDENTIFIERS National Science Foundation; *Science Achievement; *Systemic Educational Reform

ABSTRACT

During the 1990s, the National Science Foundation (NSF) funded a number of large-scale initiatives designed to change the way mathematics and science are taught in schools. These efforts, called Systemic Initiatives, shared a common emphasis on aligning all aspects of the educational system in support of a shared vision of effective mathematics and science education. In 1996, NSF provided funds to RAND to investigate the relationships between student achievement in mathematics and science and the use of instructional practices that are consistent with reform. In this report, results from the first year of the study are presented. Background information on the reforms, particularly the initiatives currently being funded by NSF, is given. Existing evidence on the effectiveness of these efforts and the difficulties researchers face in measuring relevant student outcomes is summarized. The approach to studying the problem, including the sample, measures, and methods of analysis, is described. Results from six sites are given. Finally, the conclusion of this report provides a summary of major findings, a discussion of the limitations of the analysis, and directions for future research and evaluation. Contains 31 references. (ASK)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

SE

ED 431 632

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY HAS BEEN GRANTED BY

J. Ryan

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2A

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

RAND

Teaching Practices and Student Achievement: Report of First-Year Findings from the "Mosaic" Study of Systemic Initiatives in Mathematics and Science

Stephen Klein, Laura Hamilton, Daniel McCaffrey, Brian Stecher, Abby Robyn, and Delia Burroughs

DRU-2083-EDU

May 1999

RAND Education

The RAND unrestricted draft series is intended to transmit preliminary results of RAND research. Unrestricted drafts have not been formally reviewed or edited. The views and conclusions expressed are tentative. A draft should not be cited or quoted without permission of the author, unless the preface grants such permission.

BEST COPY AVAILABLE

RAND is a nonprofit institution that helps improve policy and decisionmaking through research and analysis. RAND's publications and drafts do not necessarily reflect the opinions or policies of its research sponsors.

BEST COPY AVAILABLE

PREFACE

During the 1990s the National Science Foundation (NSF) funded a number of large-scale initiatives designed to change the way mathematics and science are taught in schools. These efforts, called Systemic Initiatives, shared a common emphasis on aligning all aspects of the educational system in support of a shared vision of effective mathematics and science education. The projects that were funded generally involved the adoption of ambitious curriculum and performance standards and the coordination of all aspects of the system in support of these goals. Particular emphasis was placed on teacher training and professional development to promote changes in instructional practice that would enable students to achieve the new standards.

Funds were given to states, to urban school districts, and to consortia of districts to implement reforms consistent with NSF's purposes. Sites had considerable flexibility in designing their programs, and they adopted very different strategies for promoting reform. As a result, initial research on the systemic initiatives focused on the complex process of development and implementation. Although individual sites gathered information about student achievement, after five years of funding NSF had no broad picture of the effects of the reform on student achievement.

In 1996, NSF provided funds to RAND to investigate the relationships between student achievement in mathematics and science and the use of instructional practices that are consistent with systemic reforms. The study was conducted in two waves: a set of six sites (including both states and urban districts) were studied during the 1996-97 school

year, and a similar set of six sites were studied during the 1997-98 school year. The same basic analytic design was replicated in each site, and the study draws much of its power and generalizability from this replication.

This report presents results for the first wave of the study, i.e., six sites that were implementing systemic reforms in 1996-97. The results should be of interest to educational policymakers at all levels of government, as well as to program developers and school administrators interested in mathematics and science education.

ACKNOWLEDGMENTS

The Mosaic project would not have been successful without the cooperation and support of key individuals in each of the participating sites. These sites included State Systemic Initiatives in Connecticut, Louisiana, South Carolina and Massachusetts, Urban Systemic Initiatives in Philadelphia, Pennsylvania; Columbus, Ohio; San Francisco, California; and Detroit, Michigan; and Local Systemic Change projects in Fresno, California; El Paso, Socorro, and Ysleta, Texas and Minneapolis, Minnesota. In each of these sites we are indebted to the leadership of the reform program as well as the regular state and district administration. A partial list of those who contributed directly to this project includes the following individuals:

Fresno, CALIFORNIA

- Charles E. McCully, Superintendent
- Georgina Takemoto, PI, Assistant Superintendent
- Robert Grobe, Administrative Analyst
- Paul Messonheimer

San Francisco, CALIFORNIA

- Dr. Waldemar Rojas, Superintendent
- Maria Santos, Associate Superintendent
- Pat Anderson

CONNECTICUT

- Richard Cole, Executive Director, Connecticut Academy for Education

- Dr. Douglas Rindone, Chief, Bureau of Research and Student Assessment, State Department of Education
- Steve Leinwand, State Mathematics Consultant
- Mari Muri, State Mathematics Consultant
- Steve Weinberg, State Science Consultant
- Dr. Bob Rosenbaum, PIMMS Department

LOUISIANA

- Cecil Picard, State Superintendent
- Kerry Davidson, Project Director, LaSIP
- John Wallin, LaSIP
- Carl Frantz, Coordinator for Evaluation, LaSIP and LaCept
- Dr. Richard Anderson, LaSIP
- Faimon Roberts, Assistant Director for Science, LaSIP
- Barbara Andrepointe, Management Information Systems, Louisiana Department of Education

MASSACHUSETTS

- Dr. David Driscoll, Deputy Commissioner, Massachusetts Department of Education
- Dr. Penny Noyce, Co-Principal Investigator Massachusetts Systemic Initiative
- Dr. Michael Silevitch, Co-Principal Investigator, Massachusetts Systemic Initiative

- Dr. Ronald Latanision, Co-Principal Investigator, Massachusetts Systemic Initiative
- Thomas Noonan, Program Director, Massachusetts Systemic Initiative
- Jeff Nelhaus, Massachusetts Department of Education
- David Perda, Massachusetts Department of Education

Detroit, MICHIGAN

- Sharon Johnson-Lewis, Assistant Superintendent, Research and Evaluation
- Mrs. Juanita Chambers, Director, Office of Mathematics and Science
- Linda Ledick
- Harvey Czuwinski

Minneapolis, MINNESOTA

- Patricia Hoben, Assistant Director, The Bakken Library and Museum
- Ken Jeddeloh, Teacher and Instructional Services, Minneapolis Public Schools
- Char Myers, Assessment, Minneapolis Public Schools
- Carol Johnson, Superintendent, Minneapolis Public Schools

Columbus, OHIO

- Dr. Rosa Smith, Superintendent, Columbus Public Schools
- Camille Nasbe, Columbus Public Schools
- Larry Sullivan, National Association of School Psychologists
- Marc Foors

Philadelphia, PENNSYLVANIA

- Ms. Rita Rice, Co-Director, Philadelphia Urban Systemic Initiative

- Mrs. Clara Tolbert, Co-Director, Philadelphia Urban Systemic Initiative
- Katherine Connor, Associate Superintendent, Philadelphia Public Schools
- Thomas Clark, Administrator, Office of Accountability and Assessment, Philadelphia Public Schools

SOUTH CAROLINA

- Dr. Susan Agruso, Director of Assessment, South Carolina State Department of Education
- Marc Drews, Coordinator, State Systemic Initiative
- Jim Casteel
- John Finch

TEXAS

- Kenneth George, Interim Superintendent, El Paso Public Schools
- H. Susan Schneider, Evaluation; Anna Bone, Research; and Dr. Ray Reynosa, USI, El Paso Public Schools
- Don Schulte, Superintendent, Socorro Independent School District
- Joe Bob Shook, Tass; and Joyce Zarowsk, USI, Socorro Independent School District
- Irma Trujillo, Interim Superintendent, Ysleta Independent School District
- Dr. Gary Ivory, Research and Evaluation; Maria Gutierrez, USI, Ysleta Independent School District
- Dr. Susana Navarro, Executive Director, El Paso Collaborative for Academic Excellence

° Gabriel Della-Piana, Director of Evaluation, El Paso Collaborative for Academic Excellence

We also wish to thank the hundreds of teachers, principals, and support staff who participated in the study, and the thousands of students who completed our assessments, without whom this complex project could never have been completed.

Horizon Research, Inc. was responsible for the design, production and scoring of the teacher surveys. We are grateful to Iris Weiss, Jon Supovitz and the staff of Horizon for their careful and competent work. CRB Associates provided data collection support in Connecticut, and Dr. Charles and Theresa Brunkerhoff deserve credit for their efforts on behalf of the Mosiac project. We also want to thank our consultants in the field who helped with coordination and data collection, including Gloria Jackson, Site Coordinator, Detroit, Michigan; Eva Nielsen, Sally Jorgensen, and Wayne Odell, Site Coordinators, Minneapolis, Minnesota; Aaron Levin and Annie Caldwell, Site Coordinators, Philadelphia, Pennsylvania; Cathe Lester, Site Coordinator, El Paso, Texas; Carmelo Sgarlato, Site Coordinator, San Francisco; Maxine Smith, Site Coordinator, Columbus, Ohio; and Donna Darby, Site Coordinator, South Carolina.

Special thanks go to Janice Earle, NSF, for her encouragement, constant support, and timely assistance. There would be no Mosaic project without her foresight and guidance.

A project of this size and complexity requires substantial operational support, and we were fortunate to have the analytical support of Tor Ormseth and the logistical

assistance of Robert Reichardt and Peter Scott, all of RAND. Helen Rhodes and Donna White contributed to the preparation of this report.

TABLE OF CONTENTS

PREFACE iii

ACKNOWLEDGMENTS v

INTRODUCTION 1

- The Systemic Initiatives (SI) Programs 2
- Earlier Evaluations of Systemic Initiatives 4
- Evidence of Relationships Between Teaching Practices and Achievement 6
- Measuring Student Achievement 7
- Overview of the Mosaic Project 8

METHODS 11

- Site Selection 11
- Data Collection 13
- Participation Rates 18
- Analysis 19

SUMMARY OF FIRST-YEAR RESULTS 20

- Distributions of Teaching Practices 20
- Relationships Between Teaching Practices and Student Achievement 23
- Pooled Analysis of Standardized Regression Coefficients 30
- Differences Between Test Formats 33

DISCUSSION 36

- Summary of Year 1 Findings 36
- Study Limitations 38
- Plans for Future Data Collection and Analysis 40

APPENDIX A 42

APPENDIX B 49

APPENDIX C 53

APPENDIX D 78

APPENDIX E.....80
REFERENCES.....82

LIST OF TABLES

1. Details on Year 1 Participating Sites 13

2. Sites, Subjects, and Assessment Instruments. 15

B1. Items on Reform Practices Scale for Mathematics 49

B2. Items on Traditional Practices Scale for Mathematics 50

B3. Items on Reform Practices Scale for Science 51

B4. Items on Traditional Practices Scale for Science 52

C1. Regression Models for State Open-Ended Mathematics Test (Grade 4), Site 1 53

C2. Regression Models for State Grid-In Mathematics Test (Grade 4), Site 1 54

C3. Regression Models for State Multiple-Choice Mathematics Test (Grade 4), Site 1 55

C4. Regression Models for RAND Hands-On Science Tasks (Grade 5), Site 2 56

C5. Regression Models for SAT-9 Multiple-choice Science Test (Grade 5), Site 2 57

C6. Regression Models for SAT-9 Open-Ended Mathematics Test (Grade 5), Site 2 58

C7. Regression TAAS Multiple-Choice Mathematics Test (Grade 5), Site 2 59

C8. Regression Models for CSIAC Hands-On Science Test (Grade 5), Site 3 60

C9. Regression Models for CSIAC Multiple-Choice Science Test (Grade 5), Site 3 61

C10. Regression Models for SAT-9 Open-Ended Mathematics Test (Grade 5), Site 3 62

C11. Regression Models for CTBS Multiple-Choice Mathematics Test (Grade 5) Total Score, Site 3 63

C12. Regression Models for CTBS Multiple-Choice Mathematics Test (Grade 5) General Mathematics Score, Site 3 64

C13. Regression Models for CTBS Multiple-Choice Mathematics Test (Grade 5) Computation Score, Site 3 65

C14. Regression Models for CSIAC Hands-On Science Test (Grade 5), Site 4 66

C15. Regression Models for CSIAC Hands-On Science Test (Grade 5), Site 4 (Continued) 67
C16. Regression Models for CSIAC Multiple-Choice Science Test (Grade 5), Site 4 68
C17. Regression Models for CSIAC Multiple-Choice Science Test (Grade 5), Site 4 (Continued) 69
C18. Regression Models for RAND Hands-On Science Tasks (Grade 7), Site 5 70
C19. Regression Models for SAT-9 Multiple-Choice Science Tests (Grade 7), Site 5 71
C20. Regression Models for SAT-9 Open-Ended Mathematics Tests (Grade 7), Site 5 72
C21. Regression Models for LEAP Multiple-Choice Mathematics Tests (Grade 7), Site 5 73
C22. Regression Models for MAT-7 Multiple-Choice Science Test (Grade 7), Site 6 74
C23. Regression Models for MAT-7 Multiple-Choice Mathematics Test Total Scores (Grade 7), Site 6 75
C24. Regression Models for MAT-7 Mathematics Concepts Subtest (Grade 7), Site 6 76
C25. Regression Models for MAT-7 Mathematics Procedures Subtest (Grade 7), Site 6 77
D1. Results from Pooled Analyses of Relationships Between Practices and Achievement 79
E1. Standardized Coefficients for Models Predicting Differences Between Formats 80
E2. Results from Pooled Analyses of Differences Between Formats 81

LIST OF FIGURES

1. Distribution of Teacher Scores on Reform and Traditional Scales by Site and Subject 22

2a. Estimated Coefficients by Site and Pooled Across Sites for Mathematics, Reform Practices, Open Ended Tests 24

2b. Estimated Coefficients by Site and Pooled Across Sites for Mathematics, Reform Practices, Multiple Choice Tests 24

3a. Estimated Coefficients by Site and Pooled Across Sites for Science, Reform Practices, Open Ended Tests 25

3b. Estimated Coefficients by Site and Pooled Across Sites for Science, Reform Practices, Multiple Choice Tests 25

4a. Estimated Coefficients by Site and Pooled Across Sites for Mathematics, Traditional Practices, Open Ended Tests 26

4b. Estimated Coefficients by Site and Pooled Across Sites for Mathematics, Traditional Practices, Multiple Choice Tests 26

5a. Estimated Coefficients by Site and Pooled Across Sites for Science, Traditional Practices, Open Ended Tests 27

5b. Estimated Coefficients by Site and Pooled Across Sites for Science, Traditional Practices, Multiple Choice Tests 27

INTRODUCTION

Many of the educational reforms that are currently underway seek to improve mathematics and science achievement by fostering classroom practices designed to enhance the development of critical thinking and problem solving skills, particularly among traditionally underserved groups of students. One approach being widely implemented today is called "systemic reform" because it attempts to align all parts of the educational system—curriculum, instruction, assessment, teacher preparation, and state and local policies such as graduation requirements—with the goal of promoting change in the classroom and, ultimately, improving student performance (Smith & O'Day, 1991). Systemic reform efforts resulted in part from observations that addressing one component of the educational system tended to be ineffective due to constraints imposed by other parts of the system (Hill, 1995; Knapp, 1997).

In this report, we present results from the first year of a study that was designed to investigate relationships between student achievement in mathematics and science and the use of instructional practices that are consistent with systemic reforms. We begin with background information on the reforms, particularly the initiatives currently being funded by the National Science Foundation. We summarize existing evidence on the effectiveness of these efforts and the difficulties researchers face in measuring relevant student outcomes. We then describe our approach to studying the problem, including our samples, measures, and methods of analysis. Following that, we present the results from the six sites where we collected data during the first year of the project. The conclusion

of this report provides a summary of our major findings, a discussion of the limitations of the analysis, and directions for future research and evaluation.

THE SYSTEMIC INITIATIVES (SI) PROGRAMS

In 1990, the National Science Foundation (NSF) launched a series of initiatives designed to promote standards-based systemic reform of mathematics and science education. Through its Statewide Systemic Initiatives (SSI) program, NSF awarded grants to 25 states and the Commonwealth of Puerto Rico from 1991 to 1993. The state level was chosen as the initial target in part because NSF viewed state policymakers as being uniquely able to influence all aspects of the educational system, including teacher training in institutions of higher education. Grants were awarded for a 5-year period, but some states were able to renew their grants for additional years.

The Urban Systemic Initiatives (USI) program, established in 1993, targets cities where large numbers of children live in poverty. This program has funded 20 large urban districts with awards of up to $15 million over five years. The program is described as a "comprehensive and systemic effort to stimulate fundamental, sweeping, and sustained improvement in the quality and level of K-12 science, mathematics, and technology (SMT) education" (Williams, 1998, p. 7). The Local Systemic Change (LSC) Program was created in 1995 to fund district-based teacher enhancement through curriculum implementation in over 50 sites. These projects are also of five years duration, but typically smaller in scope with funding based on the number of teachers served. Most of the projects are receiving funds of $2 million to $6 million over the five-year funding period. A Rural Systemic Initiatives (RSI) program operating in several sites completes

the set of NSF initiatives. Taken together, these Systemic Initiative (SI) programs have received approximately $100 million per year in NSF funding. In addition, most sites supplemented their NSF grants with local contributions. For example, sites are currently using Title I funds, corporate donations, and grants from private foundations to support and expand their systemic initiatives (Williams, 1998).

Although these programs vary in scope and emphasis, all are relatively long-term (five years, with a small number of SSIs being extended for an additional five years), and all attempt coordinated reform, aligning various parts of the educational system with one another. These initiatives, in theory at least, generally involve the development of ambitious curriculum and performance standards and the mobilization of all components of the system to support and enable all student to reach those standards (Consortium for Policy Research in Education, 1995a).

To be effective, these reforms must ultimately be adopted by teachers and take hold in the classroom (Tyack & Cuban, 1995). Thus, a primary emphasis of the systemic reform initiatives involves promotion of teaching practices that are assumed to facilitate student learning. Most initiatives offer professional development to teachers, and this component constitutes a fairly large proportion of the budget. For example, the SSI sites spent nearly one third of their first-year budgets on in-service training for teachers, more than on any other category of spending (Shields, Corcoran, & Zucker, 1994). The goal of most of this training is to increase teachers' use of classroom practices that are believed to improve achievement.

The kinds of practices being promoted by NSF, and by numerous other agencies and reformers, are based on standards documents such as those published by the National Research Council (1996), the American Association for the Advancement of Science (1993), and the National Council of Teachers of Mathematics (1989). Common to all of these documents is an emphasis on instruction that engages students as active participants in their own learning and that enhances the development of complex cognitive skills and processes. Specific practices that are endorsed include cooperative learning groups, inquiry-based activities, use of materials and manipulatives, and open-ended assessment techniques.

EARLIER EVALUATIONS OF SYSTEMIC INITIATIVES

Numerous evaluations have been conducted by the individual SI sites and by outside organizations. Most of these evaluations have focused on the degree of implementation of the reform (e.g., type and frequency of professional development offered to teachers, level of participation among teachers) rather than on student outcomes. However, NSF and the sites are becoming increasingly concerned about student achievement. Many of the SI sites have reported improvement in student test scores (Williams, 1998), but most offer little if any evidence to tie this improvement directly to SI participation.

A large-scale study of SSI programs conducted by SRI International revealed small but statistically significant differences in test scores that favored participating over non-participating schools in four of seven sites (Laguarda, 1998). However, this study had a number of serious limitations. First, the analyses did not control for any pre-existing

differences in the teachers and students found in SSI and non-SSI schools. We have observed that sites often implement large-scale reforms in phases, and those schools that participate in the earlier phases differ in important ways (e.g., the experience, interest and enthusiasm of teachers; the socioeconomic background of students) from those that participate in later phases. Second, the analysis did not examine the degree of implementation of the reform within schools. The fact that a school is considered part of the reform effort does not guarantee that all the teachers in the school are responding in the intended manner. Other researchers have found that teachers' use of reform practices is influenced by many factors, including the nature and frequency of professional development participation (Cohen & Hill, 1998; Weiss, Montgomery, Ridgway, & Bond, 1998) and the degree to which they understand the subject matter (Cohen & Ball, 1990). Third, the data were collected and analyzed by site personnel rather than by the external evaluators, and no effort was made to address differences in the quality of procedures across sites.

The absence of good evaluation data on SI programs has led some policymakers to express skepticism about the value of these programs (Fox, 1998). Others have called for more rigorous evaluations that focus on student achievement and relate it to the degree of implementation of the reform. There is some evidence of a positive relationship between the practices promoted by the systemic initiatives and student achievement in mathematics and science.

EVIDENCE OF RELATIONSHIPS BETWEEN TEACHING PRACTICES AND ACHIEVEMENT

If the systemic initiatives do result in improved student achievement, it is undoubtedly due in large part to what occurs in the classroom. For this reason, professional development and the promotion of good instructional practices are critical to the success of the initiatives. Recent as well as earlier research has provided some evidence of the effectiveness of some of the individual practices endorsed by the reforms. An experiment conducted by Ginsburg-Block and Fantuzzo (1998), for example, showed that low-achieving elementary students who were assigned to problem solving or peer collaboration conditions obtained higher mathematics scores and reported higher levels of motivation than did students who received neither of these interventions. Several other studies have demonstrated the value of peer tutoring and collaboration (e.g., Fantuzzo, King, & Heller, 1992; Greenwood, Carta, & Hall, 1988; Webb & Palincsar, 1996), as well as the benefits of contextualizing instruction in real-world problems (Verschaffel & De Corte, 1997).

A few studies have focused on relationships between student achievement and teachers' use of combinations of these practices. Cohen and Hill (1998) studied teacher-reported use of several practices consistent with the 1992 California Mathematics Framework and found that frequency of use was positively related to scores on the California Learning Assessment System (CLAS) mathematics test at the school level, after controlling for demographic characteristics. The set of teaching practices examined in this study was similar to those being advocated and supported by the systemic

initiatives. Mayer (1998) found small positive or null relationships between a similar set of practices and student scores on a standardized multiple-choice test. Thus there is some evidence that, in certain contexts at least, use of reform practices is related to higher student achievement.

MEASURING STUDENT ACHIEVEMENT

One difficulty in conducting evaluations of ongoing programs in general, and of the systemic initiatives in particular, is a lack of appropriate measures of student achievement. Most states do not currently administer tests that are well aligned with the systemic reforms (Consortium for Policy Research in Education, 1995b). Part of this misalignment may arise because many large-scale tests (whether developed by the state or by commercial publishers) rely on multiple-choice items, a format that does not always lend itself to measuring many of the scientific inquiry and mathematical problem solving skills encouraged by the systemic initiatives. In addition, many state testing programs predate both the systemic reforms and the current national standards. As of 1995, twenty-one states did not test students in science at all (Bond, Braskamp, and van der Ploeg, 1996).

An additional problem with state testing programs is that most do not provide data that can be used to track progress over time. Students are only tested in selected grade levels (e.g., 4th, 8th, and 10th). Changes in scores of successive cohorts of students confound the effects of reforms with differences among the groups of students. In addition, improvements in scores over time, which are often cited as evidence of beneficial effects of reforms on student learning, may in many cases reflect inappropriate narrowing

of the curriculum or teaching to the test (Koretz & Barron, 1998; Koretz, Linn, Dunbar, & Shepard, 1991). This problem is especially likely when the tests are part of a high-stakes accountability system or when the same form of a test is administered multiple times.

OVERVIEW OF THE MOSAIC PROJECT

The purpose of the Mosaic project is to examine the relationship between teaching practices and student achievement in mathematics and science, a relationship which is at the heart of the systemic initiatives. Our approach is to model this relationship rather than to compare directly the performance of students whose teachers participated in different phases of the reform. One of the advantages of this approach is that we can include measures of student demographic characteristics as well as prior achievement in the model. This effectively controls for differences in the students served by different teachers. Another advantage is that we measure directly the degree to which teachers actually use both traditional and reform practices, and we focus on instruction at the classroom level rather than the school level. Although in many sites the school is the unit of participation in the reform, we have found that there is substantial variation in teaching practices among teachers within a school, even though teachers may have been exposed to the same training. By collecting data on individual teachers we can address these differences in our analysis. Finally, we are measuring student achievement using both multiple-choice and open-response tests, including some hands-on science tasks that we developed and administered ourselves. This provides greater sensitivity to potential gains in skills than would be provided by multiple-choice tests alone. The Mosaic study

is being conducted in eleven sites, which will provide a strong test of the strength of the relationship across sites.

We have adopted this approach for a number of reasons. First, it is difficult to judge the effectiveness of the comprehensive systemic initiatives directly because the reforms were already well underway when this study began. It was impossible to collect baseline data and other information that would be necessary to judge the cumulative impact on of the reforms on student achievement. Second, the reforms were not implemented with an outcome analysis in mind, and in general, the sites did not address research design issues when they developed their programs. For example, some sites provided training to all teachers the first year, leaving no untreated classes to use for comparative analyses. Other sites that implemented reforms in phases defined the phases based on geographic region. As a result, student demographic factors were not the same for each wave of the reform and direct comparisons between phases would not be appropriate. In addition, few sites collected any measures of teaching, so it was impossible to know whether the training actually led to differences in classroom practices.

It is important to understand that this project is not an evaluation of the systemic reform initiatives. These initiatives are multi-faceted, multi-year efforts to bring about changes in classroom practice and in other aspects of the educational system. The reform sites have adopted a wide range of strategies to recruit and train teachers in new methods, to implement new curricula, to provide appropriate materials, to encourage and sustain change at the school level, and to instill greater interest in mathematics and science. Their success at these tasks is the subject of a comprehensive evaluation being undertaken by

SRI International (e.g., Corcoran, Shields, and Zucker, 1998; Shields, Marsh, & Adelman, 1998).

METHODS

We collected data from six of eleven sites in year one (the focus of this report), and have recently completed data collection in six sites in year two (one of the sites is providing data in both years of the study). Most of the data collection took place in elementary schools and middle schools because the bulk of the reform activities occurred at these grade levels. Our specific procedures for site selection, subject and grade level selection, and data collection are described in the following sections.

SITE SELECTION

Sites were selected to maximize the probability of encountering substantial numbers of teachers using reform practices. NSF proposed sites in which they had indications that reforms in science and/or mathematics instruction were occurring. They drew their recommendations from their site visits and from progress reports submitted by the grantees.

Mosaic project staff visited each proposed site to discuss the goals of the study, data collection components, the availability of data, student assessment measures, data requirements for linking teacher and student data, and local site coordination. All proposed sites except three were included in the study: One proposed site declined to participate because its program was not yet advanced enough to study. The reforms in another site were so widespread in the district that the necessary variation in teaching practice was unlikely to be found, though this site was later incorporated into a statewide site. One other site was excluded because the sample size was too small and testing was

limited. All other proposed sites agreed to participate and to provide the necessary student, teacher and demographic data. A local coordinator responsible for testing arrangements was designated by the site's administrators.

Our full sample included four states and seven urban centers. The sample was distributed across elementary, middle and high schools. In eight sites we focused on elementary schools, at three sites on middle schools, and at one site on high school students.[1]. In three sites we targeted only mathematics classrooms, at three sites we targeted only science classrooms, and at the other five sites we targeted both mathematics and science classrooms. In the first year, data were collected from 324 mathematics teachers at 97 schools, and from 303 science teachers at 103 schools. In all but one site, students received both multiple-choice and open-response and/or hands-on tests in the targeted subject(s). In one site, multiple-choice assessments were administered, but we were unable to schedule any open-response assessments.

School, Subject, and Grade Level Selection

School district and program staff at each site specified the grade level(s) and subject(s) in which they believed reform practices were most pervasive, and then nominated schools to participate in the study. The same basic research design was used at each site. We asked local staff to select approximately ten schools in which there was good reason to believe mathematics and/or science reforms had been implemented, and ten demographically similar schools in which the reforms had yet to be implemented. We

---

[1]One site included both elementary and middle school students.

used the nominations only to ensure variation in teaching practices; we did not compare the high- and low-implementing schools with one another directly. Table 1 lists the grade(s) and subject(s) of data collection and the numbers of schools, teachers, and students participating at each site during the first year of the study.

Table 1.

Details on Year 1 Participating Sites

Site	Grade	Subject	Number of schools[a]	Number of teachers	Number of students (varied by test in some sites)
1	3rd	Mathematics	17	46	804
2	5th	Mathematics	20	100	1530-1579
2	5th	Science	20	99	1506-1535
3	5th	Mathematics	18	73	1366-1451
3	5th	Science	20	74	1366-1438
4	5th	Science	19	45	909-933
5	7th	Mathematics	17	48	2667-2767
5	7th	Science	19	33	1926-1942
6	7th	Mathematics	25	57	3237
6	7th	Science	25	52	3279

[a]There was overlap between the mathematics and science schools at each site.

DATA COLLECTION

We collected three types of data at each site: student achievement test scores, teacher questionnaire responses, and student demographics. Data were collected from the six first-year sites in the spring of 1997.

Student Achievement Data

We obtained student scores on the mathematics and science assessments regularly administered at each site, and supplemented these with additional assessments, where feasible, to provide both multiple-choice and open-response scores. Supplementary tests were chosen in consultation with local staff, who were encouraged to select measures that they believe were reasonably well aligned with their reform efforts. Mosaic project staff trained exercise administrators to administer some of the supplementary measures, including RAND's hands-on tasks. All other tests were administered by the classroom teachers or by test administrators who worked at the local sites. Table 2 indicates the types of tests administered in each site.

In all but one site, students were given a standardized multiple-choice assessment in mathematics and/or science depending on the site designation, and an open-response test that required students to produce, rather than select, their responses. One site administered only multiple-choice tests, and we were unable to schedule additional testing due to time constraints. We used existing tests wherever possible, including state-developed tests and commercially available standardized tests. The column "Added for Mosaic Study?" in Table 2 indicates whether we supplemented the district or state's testing program with additional measures.

Teacher Questionnaires

Our primary measure of teaching practices in each site was a questionnaire developed by Horizon Research, Inc. (HRI). This instrument is a modified version of a questionnaire that has been used extensively by HRI to evaluate the implementation of

Table 2.

Sites, Subjects, and Assessment Instruments.

Site	Grade	Subject(s)	Tests	Format[a]	Added for Mosaic Study?
1	3[b]	Mathematics	State[c]	MC	No
			State	OR	No
2	5	Mathematics	State[c]	MC	No
			Stanford 9[d]	OR	Yes
		Science	Stanford 9[d]	MC	Yes
			RAND Levers & Friction[e]	OR (hands-on)	Yes
3	5	Mathematics	CTBS[f]	MC	No
			Stanford 9[d]	OR	Yes
		Science	CSIAC[g]	MC	No
			CSIAC[g]	OR (hands-on)	No
4	5	Science	CSIAC[g]	MC	No
			CSIAC[g]	OR (hands-on)	No
5	7	Mathematics	State[c]	MC	No
			Stanford 9[d]	OR	Yes
		Science	Stanford 9[d]	MC	Yes
			RAND Levers & Classification[e]	OR (hands-on)	Yes
6	7	Mathematics	MAT7[h]	MC	No
		Science	MAT7[h]	MC	No

[a]MC=multiple-choice, OR=open-response

[b]At this site, we studied teaching practices for third-grade teachers and measured the relationships with student test scores that were gathered during the following fall when students had advanced to the fourth grade.

[c]Refers to tests developed by the state

[d]Stanford Achievement Test Series, Ninth Edition, published by Harcourt-Brace Educational Measurement

[e]See Stecher & Klein, 1996 for a description of tasks and scoring guides.

[f]Comprehensive Test of Basic Skills, published by CTB/McGraw-Hill

[g]California Systemic Initiatives Assessment Collaborative. This test was developed by a consortium of educators and researchers, and was designed to be aligned with NSF-supported reform efforts.

[h]Metropolitan Achievement Tests, Seventh Edition, published by Harcourt-Brace Educational Measurement. In this site, we were unable to schedule any open-response testing.

systemic reform initiatives. Questionnaires were administered to all teachers in a school teaching the targeted subject and grade level. Typically, the site coordinator or assistant distributed the questionnaires either individually or at after-school meetings and then collected completed questionnaires in individual sealed envelopes for return to RAND.

We created separate questionnaires for mathematics and science teachers, but many of the items were identical across subjects. Questions asked teachers to report the frequency of various instructional practices ranging from traditional (e.g., "have students watch me [teacher] do a science demonstration") to reform ("[students] conduct investigations where they develop their own procedures for addressing a question or problem"). General topics included: amount of time spent on science/mathematics; approach to introducing a new topic; typical teacher instructional practices; typical student activities; types of written assignments; teacher use of students' written work; and methods of assessing student learning.

In addition, teachers were asked to complete a brief demographic section, providing information about degree, certification, coursework in mathematics and/or science, gender, ethnicity, and years of teaching experience. In sites where science or mathematics specialists delivered instruction instead of the regular classroom teacher, we also asked the respondent to identify the teaching situation.

HRI, acting as a subcontractor to RAND, developed the questionnaire, processed the data, and prepared an analysis file for RAND. HRI also validated the instruments in one of the six first year sites. In this site RAND selected a sample of schools where teachers were expected to have a wide range of implementation of the reform. A local

coordinator scheduled interviews with 40 teachers from these schools. Where the schedule permitted, the coordinator also arranged for the teachers to be observed teaching mathematics or science. Trained staff from HRI interviewed each teacher about a recent science or mathematics unit. Teachers were asked to discuss their goals for the unit, the extent to which students engaged in investigations and collaborations, and the types of assessment they used. Teachers were also asked to show associated artifacts (student work, journals, assignments, etc.).

On the basis of the interviews, artifacts, and observations, the validator rated the instructional program of each teacher on a five-point scale on each of three dimensions: the use of student-centered strategies, the investigative culture of the classroom, and the use of reform-oriented strategies. The validator also made an overall judgment of the degree to which each teacher's instructional program reflected reform practice. These ratings were compared with the self-reported practice information from the teacher surveys. The analysis showed a positive relationship between the composite survey scale and validators' overall ratings. This relationship was stronger in mathematics than in science, but it was positive for both subjects. The results of this analysis support the validity of the surveys as measures of teaching practices.

Demographic Data

Information about student characteristics was obtained from the sites for three purposes. First, these data were used to verify that comparison schools were similar to implementing schools on student demographics, enrollment, and grade span. Second, the data were included as covariates in the analysis of relationships between teaching

practices and student achievement. Finally, they enabled us to study whether these relationships varied as a function of student characteristics. In most sites we obtained data on students' race/ethnicity, gender, participation in free or reduced-price lunch programs, language background, and participation in special education or gifted programs. In many sites we also obtained student test scores from the previous year.

PARTICIPATION RATES

Participation in the study by schools, teachers and students was quite high. It is difficult to summarize participation in a simple manner because of the multi-stage nomination and enrollment process, differences in procedures across sites, and partial participation by some schools, teachers and students. Nevertheless, across the six sites, between 85 percent and 100 percent of the schools we initially contacted participated in the study, i.e., tests and teacher surveys were administered and completed by the majority of eligible individuals. Similarly at the site level, between 71 percent and 98 percent of the teachers who received a survey completed it. Student participation rates were more difficult to compute because of partial participation and because of a variety of testing exclusions. In addition, we could not always link students and teachers. Nevertheless, across the six sites, between 65 percent and 94 percent of the students identified as being taught by the teachers in the study completed all the desired tests in mathematics or science. Detailed descriptions of school, teacher and student participation in each of the sites will be found in Appendix A.

ANALYSIS

The primary purpose of this study was to investigate the degree to which student achievement was associated with teachers' use of instructional practices consistent with the reforms. We explored these relationships using linear regression analysis, which enabled us to control for student background characteristics and previous test scores. At each site, we conducted separate analyses for mathematics and science, for open-response and multiple-choice tests, and for reform and traditional practices. We fit these models using individual student data, with all students from the same classroom receiving the same values for the teacher reform and traditional scales, and we used an adjusted standard error estimate to account for possible correlation among responses from students with the same teacher (McCaffrey & Bell, 1997).

SUMMARY OF FIRST-YEAR RESULTS

In this section we present our results from the six sites where we collected data during the first year of the project. We first present summaries of teachers' reported use of reform and traditional practices. We then present our findings with regard to the relationships between use of these practices and student achievement in each site. Finally, we describe results of an analysis of differences between open-response and multiple-choice achievement measures.

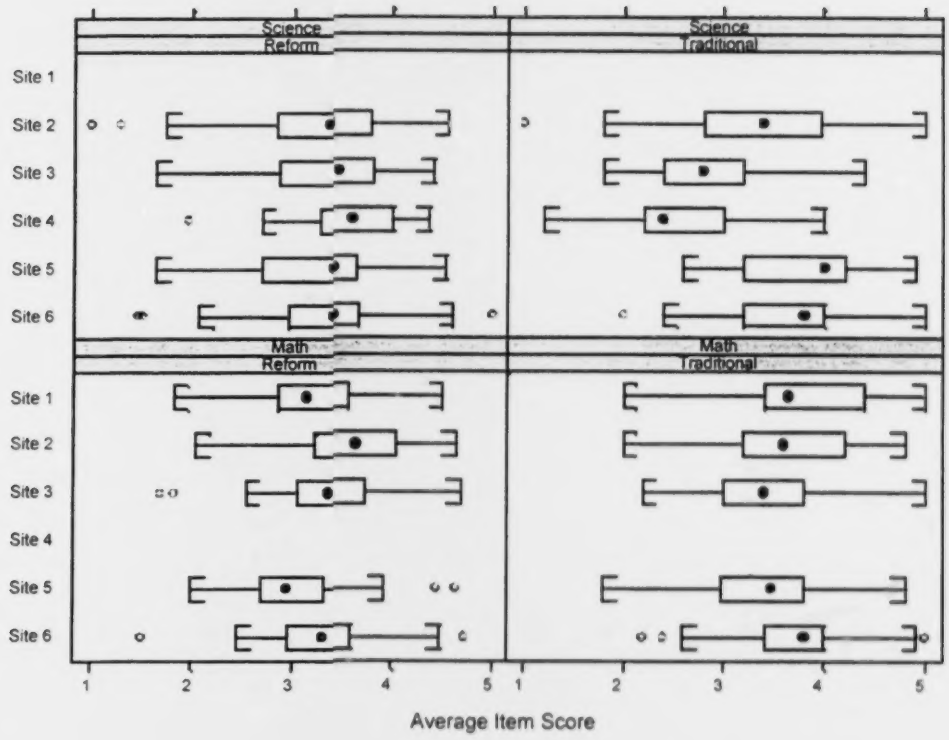
DISTRIBUTIONS OF TEACHING PRACTICES

Based on exploratory factor analyses of the teacher questionnaire items, we identified two clusters of items and created scales from these by simply summing the scores on each item. The first scale measured the teacher's use of "reform practices." Each of the 22 items in this scale asked teachers to report the frequency of use of a particular reform practice (such as cooperative groups, portfolios, and extended investigations). We also created a 5-item "traditional practices" scale based on items that measured the amount of time teachers spent in traditional teaching practices (such as textbook work, lectures, and short-answer tests). Appendix B provides a list of items in each scale. This distinction between reform-related practices and more traditional practices, which emerged from factor analyses conducted on each site's data, is consistent with the kinds of definitions used in other research on mathematics and science reform (e.g., Cohen & Hill, 1998). However, it is important to note that the two scales are not opposites of one another. Teachers may use both reform and traditional practices to

different degrees. Correlations between the two scales ranged across sites from moderately negative to moderately positive, with many close to zero, i.e., teachers' use of one type of instructional practice (reform or traditional) does not necessarily indicate whether they will use another type of instructional practice (reform or traditional).

In each site we observed a wide range of practices on both the reform and the traditional scales. The box and whiskers plots in Figure 1 show the distributions of mean scale scores for each combination of site and subject (mathematics or science). For these plots, the score for each teacher simply averaged the item responses across items. All items used a 5-point Likert scale, so teachers' scores could range from 1 (rarely or never using any of the practices) to 5 (engaging in all practices daily or almost daily). The solid dot indicates the average score for all teachers. The lower end of the box is the 25th percentile of the distribution and the upper end of the box is the 75th percentile, and the whiskers show the 5th and 95th percentiles. Outlier values are shown as individual points in the plots.

Overall, teachers report practices similar across the sites, with the most variation occurring for traditional practices in science. For example, science teachers' average scores on the reform scale ranged from 3.27 in Site 2 to 3.57 in Site 4. Site averages on the reform scale for mathematics teachers were somewhat more variable, ranging from 3.01 to 3.61. On the traditional practices scale, site averages for science were more variable than averages for mathematics. The former ranged from 2.65 to 3.78, whereas the latter ranged from 3.33 to 3.73. Interestingly, sites where teachers' average use of reform practices were quite similar showed fairly large discrepancies in their use of traditional practices.



Note: Box and whisker plots show mean score (dot) 25th to 50th percentile range (rectangle or box), and 10th to 90th percentile range (whiskers). Outliers appear as open circles.

Figure 1.
Distribution of Teacher Scores on Reform and Traditional Scales by Site and Subject

We also found substantial variation in teaching practices within schools, regardless of degree of participation in the reform programs. Clearly some teachers in participating schools had not adopted many of the reform practices that were emphasized by the SI, whereas some teachers in non-participating schools were using these practices even though they had not been exposed to SI-specific professional development. This underscores the importance of using classroom-level reports of teaching practices rather than simply studying differences at the school level.

RELATIONSHIPS BETWEEN TEACHING PRACTICES AND STUDENT ACHIEVEMENT

In this section we describe the results of our analyses relating teaching practices to student achievement. In general, we found that teachers' use of reform practices was positively related to student achievement in most sites, but the effects were quite small. We observed slightly stronger relationships for open-response than for multiple-choice tests in all mathematics sites and in most science sites. Use of traditional practices, by contrast, was often negatively related to student achievement, particularly in mathematics, but again the relationships were weak.

Figures 2-5 provide an overview of our findings in the six first-year sites. The full regression models appear in Appendix C. The relationships reported in the figures are the estimated coefficients from our regression models for the reform and traditional practices scales. We standardized test scores and teaching practice scales so that the reported coefficient is the expected difference in test score standard deviation units for a one standard deviation unit increase in scores on the reform or traditional scale. The dark dot represents the point estimate for the coefficient and the gray bar represents 95 percent confidence interval for that point estimate. When the bar does not meet the zero line it means that the coefficient is statistically different from zero. The lowest bar in each figure shows the average coefficient from the pooled analysis, described later.

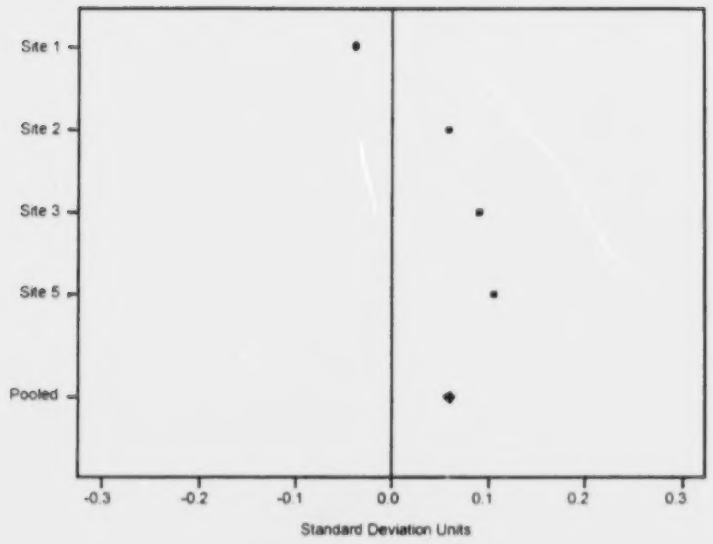


Figure 2a.
Estimated Coefficients by Site and Pooled Across Sites for Mathematics, Reform Practices, Open Ended Tests

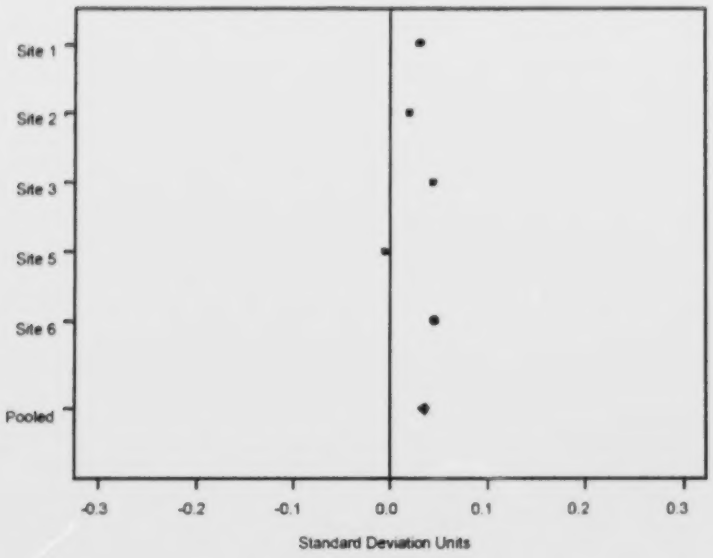


Figure 2b.
Estimated Coefficients by Site and Pooled Across Sites for Mathematics, Reform Practices, Multiple Choice Tests

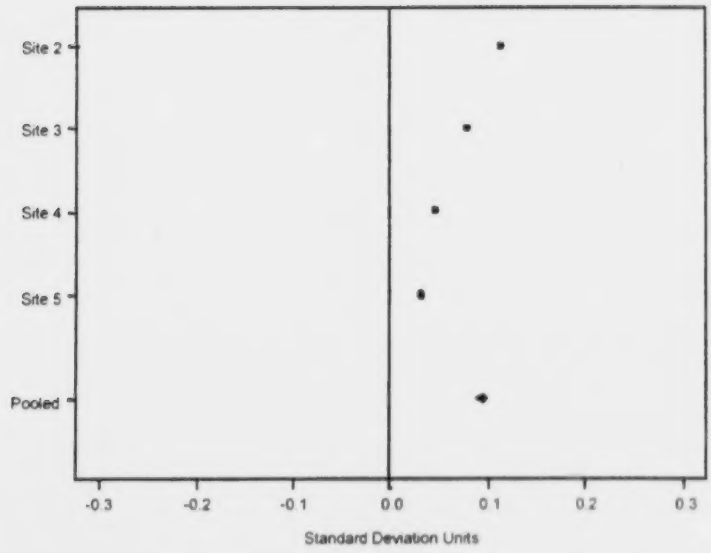


Figure 3a.
Estimated Coefficients by Site and Pooled Across Sites for Science, Reform Practices, Open Ended Tests

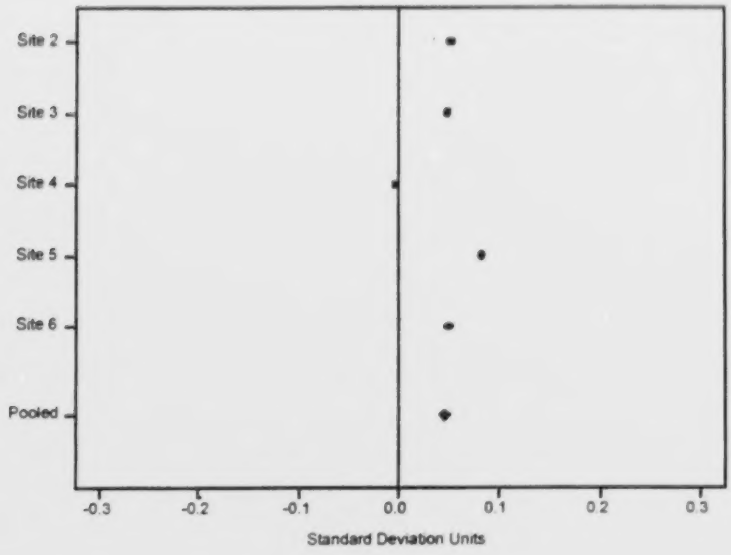


Figure 3b.
Estimated Coefficients by Site and Pooled Across Sites for Science, Reform Practices, Multiple Choice Tests

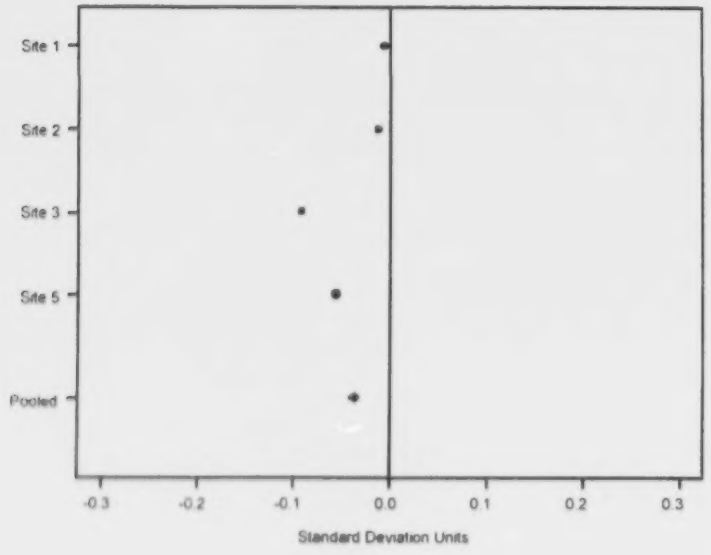


Figure 4a.
Estimated Coefficients by Site and Pooled Across Sites for Mathematics, Traditional Practices, Open Ended Tests

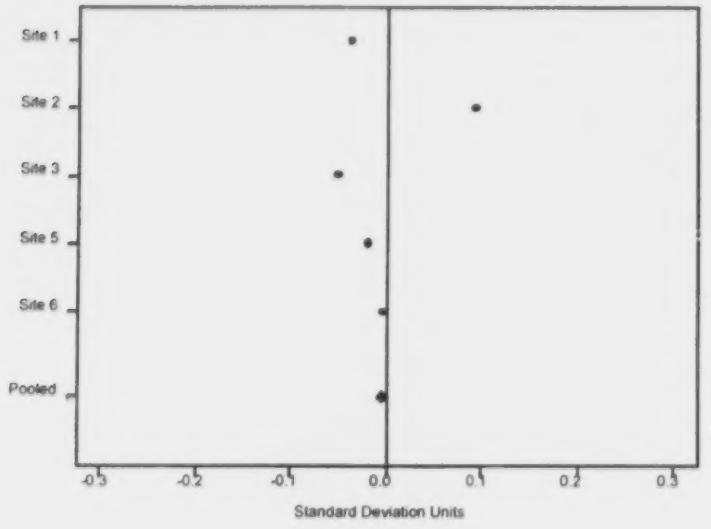


Figure 4b.
Estimated Coefficients by Site and Pooled Across Sites for Mathematics, Traditional Practices, Multiple Choice Tests

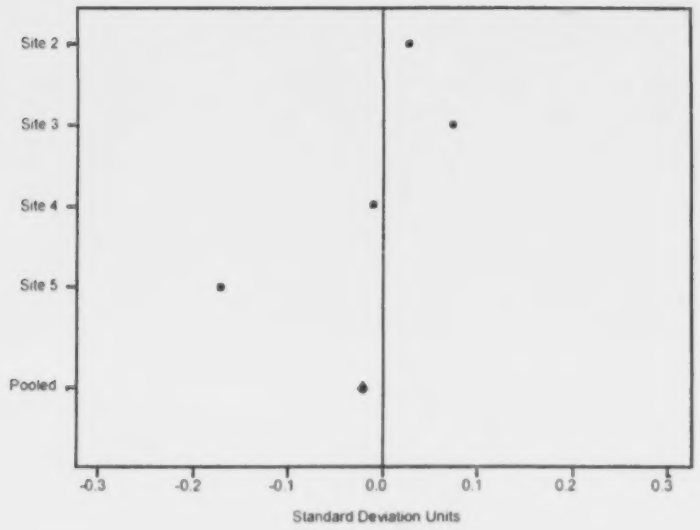


Figure 5a.
Estimated Coefficients by Site and Pooled Across Sites for Science, Traditional Practices, Open Ended Tests

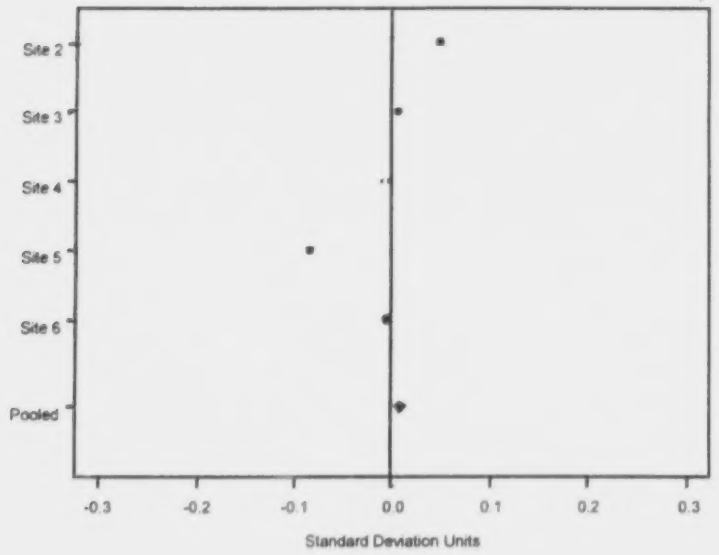


Figure 5b.
Estimated Coefficients by Site and Pooled Across Sites for Science, Traditional Practices, Multiple Choice Tests

Figure 2a, which shows relationships between use of reform practices and achievement on open-response mathematics tests, indicates that in four of the five sites where we had open-response mathematics tests, higher scores were associated with greater use of reform practices. Similarly, Figure 2b shows that for almost all of the participating sites, higher multiple-choice test scores in mathematics were also associated with greater use of reform practices. Figures 3a and 3b show that greater use of reform practices in science was associated with higher test scores on both open-response and multiple-choice measures in science.

In contrast, the majority of the relationships between use of traditional practices and student achievement were negative. For example, Figure 4a indicates that for all of the participating sites, greater use of traditional teaching practices in mathematics was associated with lower scores on open-response mathematics tests. We observed a similar pattern in Figure 4b, where for four of five sites we found greater use of traditional teaching practices associated with lower test scores on mathematics multiple-choice tests. The relationship between traditional teaching practices and science test scores was less consistent, as shown in Figures 5a and 5b. We observed two positive and two negative relationships between traditional teaching practices and open-response test scores, and two positive and three negative relationships between traditional teaching practices and multiple-choice test scores.

As shown in Figures 2-5, the relationship between teaching practices and test scores is at most small in almost all our models. For example, the largest positive relationship was between reform teaching practices and open-response science tests in

Site 2 (see Figure 3a), where the standardized regression coefficient was 0.11. In this site, our model suggests that for a teacher using all of the reform practices monthly, the average student was predicted to score at about the 48th percentile on the test, while for a teacher using all of the reform practices weekly we would predict that a similar student would score at about the 54th percentile on the test[2]. Smaller changes in percentiles would be expected in the other sites. Compared with the coefficients for most of the student background characteristics (e.g., an average coefficient of 0.54 across sites for participation in free and reduced-price lunch programs), all of the relationships we observed may be considered small.

We expected to see larger relationships between reform practices and open-response measures than between reform practices and multiple-choice measures because the former tend to be more closely aligned with the reforms. Inspection of the regression coefficients suggests that this is the case. In all sites where we had both multiple-choice and open-response scores, we found that the relationships between reform practices and open-response mathematics scores were larger than the relationships with multiple-choice mathematics scores. In science, in two out of three sites the relationships with open-response test scores were larger than the relationships with multiple-choice test scores. Later we discuss a test of the statistical significance of this difference.

---

[2]We used our model to predict the score for the "average" student (a student with all student background predictors set to the mean) with a teacher scoring 3 on each reform practices item (monthly use of reform practices). We then found the percentile of this predicted score among the test scores from the

POOLED ANALYSIS OF STANDARDIZED REGRESSION COEFFICIENTS

We conducted pooled analyses which combined data from all six sites to produce a single estimate of the coefficient relating teaching practices (reform or traditional) to student achievement. We conducted separate analyses by subject (mathematics or science), test format (multiple-choice or open-response) and teaching practice scale (reform or traditional) for a total of eight pooled analyses.

Such pooled analyses are appropriate when the coefficients from the various sites describe a single relationship. Because we examined the same relationship between teaching practices and student test scores in all six sites using the same study design and similar analyses, such a pooled analysis is appropriate. Specifically, we expect that the coefficients from our models are homogeneous and that small differences in our site studies (e.g., differences in tests and the covariates in our models) can be modeled as small random variations among the coefficients. The homogeneity assumption can be tested by examining the variability among the coefficients across the sites. Technical details on the pooled analyses and the estimates of variability are provided in Appendix C.

We conducted the pooled analyses using the estimated regression coefficients presented earlier. We did not pool the individual student data because we used different covariates in our site models so we would have had to exclude some useful predictors from the model. For example, because we did not have prior year test score for site 5,

---

site, and repeated the process for the average student with a teacher scoring 4 on each item (daily use). The

pooled analyses on the student data could not include that important predictor. However, the estimates we obtained by pooling the regression coefficients are similar to those we would have obtained by pooling individual scores and fitting a random coefficients model (Goldstein, 1995). Our approach enabled us to pool data across sites without requiring identical models in every site.

The bottom bars in Figures 2-5 show the pooled estimates of the standardized regression coefficients for each of the eight analyses. Additional detail is provided in Table D1 in Appendix D. In most cases, the variability of coefficients was sufficiently small to be within the range expected as a result of sampling error within sites. Therefore the pooled analysis is appropriate and with our small sample we have no indication that the relationship between teaching practices and student achievement is site-dependent.

One exception occurred for analysis of the relationship between reform practices and student achievement on open-response mathematics tests. The pooled coefficient was 0.06 (Figure 2a), but the test of variability indicated heterogeneity across sites. The variation was primarily a result of the negative relationship (-0.01) we observed in site 1; the other three coefficients ranged from 0.06 to 0.11. When we conducted the pooled analysis excluding site 1, the estimate was 0.09 and the test of variability indicated no heterogeneity. This difference is probably due to the fact that in site 1 we used scores on a locally developed open-response mathematics test, whereas in the other three sites we administered the Stanford 9. The relationship between instructional practices and

---

ercentile is based on our sample and is not a percentile from a national norming group.

achievement may be sensitive to the particular instrument used, or to the administration conditions. This problem is discussed further in the final section of the report.

Our pooled estimate of the relationship between reform teaching practices and student achievement on multiple-choice mathematics tests was 0.035 and was statistically significantly different from zero. For traditional teaching practices, we obtained pooled estimates that were slightly less than zero for both multiple-choice and open-response tests in mathematics, but neither was significantly different from zero. The variability among coefficients was sufficiently small to permit pooling for all of these analyses except that for traditional practices and multiple-choice tests, but again, this variability resulted primarily from a single outlier, the estimate of 0.09 from site 2.

In science, we obtained pooled estimates of 0.05 for the relationship between reform teaching practices and multiple-choice test scores, and 0.10 for the relationship between reform teaching practices and open-response test scores. Both of these estimates were statistically significantly different from zero. For traditional teaching practices in science, we obtained estimates of 0.01 for open-response scores and -0.02 for multiple-choice scores, neither of which differed significantly from zero. The latter analysis revealed relatively large variability among coefficients relating multiple-choice science scores to traditional practices, suggesting that a pooled analysis might not be appropriate in this case.

To summarize, the pooled analyses revealed small, statistically significant, positive relationships between teachers' use of reform practices and achievement on both kinds of tests and in both subjects. Results for open-response mathematics scores appear

somewhat sensitive to the particular test used, but for the other measures we detected no evidence of heterogeneity among sites in the strength of the relationships between teaching practices and achievement. In general, teachers' use of traditional practices was unrelated to student achievement.

DIFFERENCES BETWEEN TEST FORMATS

Consistent with the individual site results, the pooled analyses revealed slightly larger relationships between open-response scores and reform teaching practices than between multiple-choice scores and reform teaching practices. This finding is consistent with the hypothesis that the former type of test is more closely aligned with the reforms and therefore better able to detect effects. To test the statistical significance of this difference, we calculated the difference in standard deviation units between each student's score on the open-response test and his or her score on the multiple-choice test in the same subject. We then modeled these differences as a function of teaching practices and student background covariates. The analysis was repeated for both subjects and for all sites.

The coefficients for teaching practices obtained for each site are presented in Table E1 in Appendix E. The coefficient was positive in three out of the four sites where we studied mathematics. However, only one of the differences, 0.12 in site 5, was statistically significant. Similarly, the coefficient for teaching practices was positive in three out of the four sites where we studied science, but none of these differences was statistically significant.

We again conducted a pooled analysis of coefficients across sites. Results are given in Table E2 in Appendix E. For mathematics the pooled estimate was 0.03. This implies that across the sites the expected increase in student mathematics test scores for a unit increase in a teacher's score on the reform scale was 0.03 standard deviation units higher for open-response tests than for multiple-choice tests. However our estimate was not statistically significantly different from zero. In addition, we found a relatively large between-site variance in these estimated differences, even after controlling for sampling error within site. In other words, we found that the difference in the sensitivity of open-response and multiple-choice tests varied from site to site. In the two grade 5 sites where the open-response test was the Stanford 9, the differences were very similar, 0.048 and 0.051. In Site 5, we again administered the Stanford 9 open-response test, but to seventh graders, and the difference between open-response and multiple-choice tests was 0.119. In Site 1, we used scores from a test developed by the state and the difference was -0.041. Hence, the sensitivity of tests might depend on the both the test form and grade. Additional data are necessary to explore this hypothesis.

For science the pooled estimate was 0.04. This implies that across the sites the expected increase in student science test scores for a unit increase in a teacher's score on the reform scale was 0.04 standard deviation units higher for open-response tests than for multiple choice tests. Again, our estimate was not statistically significantly different from zero. For science there was little variability in estimated differences across sites after accounting for sampling error within sites.

Thus, although inspection of regression coefficients suggests that open-response tests function differently from multiple-choice tests, our data do not provide sufficient evidence to support the claim that the formats differ in their sensitivity to the effects of the reform. Even so, the patterns we observed indicate that there may be a difference that is worthy of further investigation.

DISCUSSION

In this final section we briefly summarize findings from the first year of the Mosaic study and discuss their implications. We also suggest possible explanations for the variability in findings across sites. We discuss several important limitations of this study, and then describe plans for future data collection and analysis.

SUMMARY OF YEAR 1 FINDINGS

As illustrated by Figures 2-5, the relationships between student achievement and teachers' use of instructional practices supported by the SI reforms tend to be positive but small, particularly in comparison to relationships between achievement and student background characteristics such as socioeconomic status and ethnicity. If in fact the observed relationships represent the effects of teaching practices on student achievement, their small magnitude may not be surprising given the brief period of time (less than one academic year) which was captured by teachers' questionnaire responses. Use of particular instructional strategies in a single course during a single school year would not be expected to lead to effects as large as those associated with student background characteristics. Several years of exposure may be needed to achieve a reasonably large effect. This suggests the need for longitudinal investigations, discussed below.

The direction of relationships was fairly consistent across sites, but their magnitudes displayed some variation. There are several potential sources of this variation. First, our models differed slightly across sites because we relied on locally available data to construct covariates. Second, various aspects of SI program

implementation, such as the amount and quality of professional development activities, undoubtedly affected the kinds of teaching practices that were used. Even if two teachers report using reform practices with similar frequency, their approaches to those practices may differ substantially and may reflect specific features of the local reform program. Third, the achievement measures used in each site varied on a number of dimensions, including psychometric quality (e.g., reliability), content, and degree of alignment with the local curriculum.

This last source of differences has implications for future evaluations of SI's and other reforms. Most evaluations rely on locally available student achievement data, in large part because it is expensive and often not feasible to administer additional measures. Many principals and teachers believe that their students spend far too much time taking the tests that are required locally, and are therefore reluctant to volunteer for supplementary testing. Locally developed and administered tests may also be preferred because they are presumed to be more closely aligned with the local reform effort than would a measure chosen and administered by outside evaluators. Although many districts and states are working to develop tests that reflect local standards and curricula, in the sites that we studied most test development lagged far behind the reform implementation, leaving local personnel to rely on tests that they did not necessarily believe were appropriate.

Our analyses revealed that in most cases, tests that we added specifically for the Mosaic study functioned better than the locally administered tests. Local tests were often less reliable than the tests we administered. In addition, local tests exhibited

unexpected relationships with other measures of student achievement and with student background characteristics that raise questions about the validity of scores. In several cases we suspect that repeated use of the same test questions in a high-stakes testing environment, where teachers may have engaged in inappropriate test preparation activities, may have lessened the quality of the scores. For example, in two sites there was an extremely high correlation between the percent of students in a school receiving free or reduced-price lunch and the school's average score on the Mosaic multiple-choice and open-ended tests, but the correlation between free or reduced lunch status and the local test (measuring similar content) was close to zero. This unexpected result may reflect efforts to "drill" students on the specific skills required by the local test.

Although the overall differences we observed between multiple-choice and open-response tests were not significant, the general pattern suggests that format effects should be investigated further. In particular, it raises questions concerning whether the two types of tests measure different constructs. Most advocates of systemic reform believe that traditional, multiple-choice tests do not adequately reflect the range of competencies that the reforms are expected to develop, and that tests requiring students to construct their answers and to engage in complex problem-solving are more appropriate. Our results do not indicate that this is necessarily the case, but the question deserves further investigation.

STUDY LIMITATIONS

There are several caveats that need to be considered when interpreting the results of this study. As with most educational research, our inability to investigate effects using

an experimental design places some limitations on the kinds of inferences that can be made from results. Perhaps the primary problem is that without random assignment of students and teachers to treatments, we cannot be certain that the relationships we observed can be attributed solely to classroom practices. There may be other differences in student characteristics across classrooms that contribute to differences in performance and that influence what teachers do. For example, teachers may tend to engage in more reform-based practices with higher achieving students. Controlling for prior achievement, as we do in most sites, is helpful but does not eliminate the problem completely.

A second limitation is the lack of information on what led teachers to utilize particular practices. Some may have adopted certain strategies as a result of participation in the professional development activities that are provided by the SI funds, but there are many other potential sources. The large variability in teaching practices within schools, which was observed for SI as well as non-SI schools, suggests that factors other than SI participation are influencing teachers' decisions about how to teach. Our initial intent was not to determine the reasons for teachers' use of practices, but information on this would be helpful to those who are designing and implementing professional development programs.

A third weakness of our approach stems from the use of questionnaires to measure instructional practices. Like any such measure, our items are subject to inaccurate responses, particularly those that reflect social desirability. Perhaps more importantly, our questions addressed only the frequency with which teachers used particular practices and did not address the way in which they were used or the overall

quality of the instruction. Clearly, some approaches to using cooperative groups are more effective and more consistent with the intent of the reform than others, but we cannot detect these differences using our questionnaires. Multiple classroom observations, interviews, and inspection of classroom materials would undoubtedly provide a better measure of instructional practice. This type of data, however, is considerably more expensive to collect, and is usually only done on a small scale.

Finally, as discussed earlier, our analyses focused on students' exposure to practices during a single academic year, and therefore do not allow us to follow the development of teacher experience in reform practices or the impact of student exposure to these practices across several years. Many of the achievement tests used at our sites require students to apply knowledge and skills they have gained over a number of years, and thus performance on these tests is undoubtedly influenced by students' instructional experiences in prior years. Project directors at some of our sites report that student exposure to reform practices is typically small, even in sites that have been grant recipients for several years. This occurs because programs generally need time to be fully implemented. It is widely believed that students must be exposed to reform practices for more than a single year before the effects of these practices on achievement become clearly evident, but there is currently little data available to support or refute this claim.

PLANS FOR FUTURE DATA COLLECTION AND ANALYSIS

During the second year of the Mosaic project, data were collected from five additional sites. The design of the data collected was essentially the same as the first year, and will provide us with additional estimates of the relationships between teaching

practices and student outcomes. During year 2 we also collected data from one of the sites that participated during the first year. The same grade level was used, so we will not be able to track students over time. However, results from this site will enable us to explore changes in teaching practices by individual teachers, and to examine what happens to student achievement as schools and teachers are involved in the reform for longer periods of time.

We have also planned a longitudinal study using one of the sites that participated during year 2. In this site, we hope to follow students over three years, collecting data on the instructional practices used by their teachers during each year. This will enable us to conduct a multi-year "dose-response" study in which degree of exposure to the practices can be related to student growth in achievement over a longer time period.

APPENDIX A

Participation of Each Site

Site 1

Our sample in site 1 includes 4th grade students from 17 schools. Our original sample consisted of 20 schools but 2 schools refused to participate, one nominated as highly involved and one nominated as a control. We also excluded a school where no grade 3 teachers responded. This school was also nominated as highly involved in the reform. From these 17 schools, we obtained survey responses from 46 of the 60 grade 3 teachers we surveyed.

We linked 1012 grade 4 students to the responding teachers. For our analyses, we eliminated students who were missing scores on the 4th grade test, as well as a handful that were missing data on other covariates. This left 804 students for our analyses.

Site 2

Our site 2 sample contains students from 20 schools. The sample included 10 schools nominated as highly engaged in the reform and 10 matched control schools. The site's systemic reform included both mathematics and science, and schools involved in one reform were expected to be involved in the other. Hence, we used the same sample of 20 schools to study the effects of both mathematics and science teaching practices.

We surveyed 115 grade 5 teachers in these 20 schools. We asked every teacher about teaching practices for both mathematics and science. We obtained responses for 100 mathematics and 99 science teachers.

There were 2345 fifth graders in these 20 schools.[3] However, we could not accurately link 45 students from one school to their science teacher because teachers at this school shared teaching responsibility in an informal manner that was not documented. We excluded these students and their teachers from our analyses. In addition we excluded 224 special education students and 427 students who did not complete the grade 5 state multiple-choice tests. We restricted our attention to the remaining students as a means of identifying fifth grade students whose scores on our English language standardized tests would not be invalidated by limited English proficiency or special needs. Of the remaining 1649 students, 1513 completed both hands-on science tasks, and 1541 completed the multiple-choice science test. Because we also exclude from our analyses a handful of students missing demographic covariates, our final analysis sample for the hands-on tests contained 1535 students and our final sample for the multiple-choice test included 1506 students.

In mathematics, our two outcome measures were the state's grade 5 multiple-choice mathematics test and a RAND administered open-ended mathematics test. For our mathematics sample, we excluded 151 students who received scores of 0 on the state's mathematics test because they appear to be outliers in a number of respects (although including them has no substantial impact on results). We also excluded special education students as well the handful of students who were missing data on any of the other

---

[3]Fifth grade classrooms are any classroom with fifth grade students. Because several schools in our site 2 sample use multi-grade classes some of our fifth grade classrooms include fourth or sixth grade students as well as fifth graders. However, only fifth grade students are included in our study.

covariates. Finally, we eliminated 45 students whose links to teachers could not be verified. This left a sample of 1579 students for the analysis of multiple-choice scores, 1530 of whom also completed the open-ended mathematics test.

Site 3

Our site 3, mathematics sample contains students from 18 schools in from across the school district. The sample included 10 schools nominated as highly engaged in the reform and 8 matched control schools. Our original sample included 20 schools, 11 nominated as highly engaged and 9 matched controls. However, we obtained no teacher survey responses from two schools and data from these schools are excluded from our analyses. Our science sample from site 3 includes 20 schools from the district; 10 nominated as highly engaged and 10 matched controls. Five schools are included in both the mathematics and the science samples.

We obtained survey responses from 73 of the 87 mathematics teachers in our mathematics sample and 74 of the 85 science teachers in our science sample. Only students whose teachers responded are eligible for our study.

There were 1498 eligible fifth graders in our mathematics sample. We treat as ineligible: 1) students whose teacher did not completed a survey; 2) students exempted from district testing because of LEP or special education status; 3) students not in grade 5, even if the students were in mixed grade level classrooms. We include in our study all students identified as LEP or special education but not exempted from district testing. We believe that if the district to uses these students' scores to measure school outcomes, then they are appropriate for measuring the effects of teaching practices. Of these 1498

students, 1366 completed the open-ended mathematics test. 1451 students completed both the general mathematics and the computation subtests of the district's multiple-choice mathematics test. An additional 16 students completed only the general mathematics portion and 6 students completed only the computation portion of the test.

We used identical criteria to create the science sample, leaving us with 1652 eligible fifth graders. Of these 1652 students in the science sample, 1438 were administered the multiple-choice science test and 1366 were administered a hands-on science test.

Site 4

Our original sample included 25 schools, 11 nominated as high implementing and 14 as controls. Our teacher rosters from the district indicated that 87 teachers taught fifth grade students in these schools. However, 13 of these teachers taught only special education students none of whom were tested by the district. We determined that these teachers were special education teachers and that they and their students were not eligible for our study. Thus, our original sample included 74 eligible teachers and 1566 eligible students from 25 schools. Of the 74 eligible teachers, 52 responded to our survey.

We dropped four schools from this sample of 25 schools because the schools conducted no science testing. Two of the schools that we dropped were nominated as high implementing and two were controls. By dropping these four schools from the sample, we lost seven eligible teachers and 140 eligible students. Hence, after excluding these schools from our sample, we have 21 participating schools with 67 eligible teachers and 1426 eligible students. Of the remaining 67 teachers, 49 (73 percent) responded to our survey.

From this sample of 21 remaining schools, we dropped two additional schools, with a total of 5 eligible teachers, because no teachers from these schools responded to our survey. Hence our final sample included 19 participating schools, 8 nominated as high implementing and 11 controls. This sample contained 62 eligible teachers and 1314 eligible students. From this sample of 62 teachers, we have a total of 49 who responded to our survey and a total 59 whose students completed science testing.

Of the 49 teachers who completed the survey, only 45 also conducted science testing. Hence our final sample consists of the 45 teachers completed our survey and conducted science testing. This sample contains1012 students. 0f these 1012 students only 954 complete any science tests: 933 complete the multiple choice test and 909 complete the hands-on tasks.

Site 5

In site 5 we started with a sample of 20 schools (10 nominated as engaged in the reform and 10 nominated as yet to be involved) for studying the effects of teaching practices for mathematics instruction and student outcomes on standardized mathematics tests. We began with a separate sample of 20 schools (10 engaged and 10 yet to be involved) for studying the effect of teaching practices for science instruction on student outcomes in science. Two schools were in both the mathematics and science samples. Also, three schools refused to participate in our mathematics study and one refused to be in our science study.

We obtained survey responses from 48 of the 49 mathematics teachers in our mathematics sample and 33 of the 46 science teachers in our science sample. Only students whose teachers responded are eligible for our study.

For our mathematics sample, there were 3199 students in the participating schools. 2940 students completed the open ended mathematics test and 3028 completed the state's multiple-choice mathematics test. Some students were not tested because of absence from school or because they are exempted from testing (e.g., special education students are sometimes excluded from testing). We excluded special education students and 48 students who received a score of zero on the open-ended test.[4] An additional 4.7 percent of the students were missing free and reduced lunch status, and a handful of records were missing other variables used in the model. We excluded these students from our analyses. Hence our final analysis samples included 2667 students with open-ended mathematics test scores and 2767 students with multiple-choice mathematics scores.

For our science sample, included 2436 students in the participating schools. 2053 students completed both the hands-on science tasks, and 2093 completed the multiple-choice science test. Again, we excluded special education students from our analysis. We also excluded an additional three percent of students whose data were missing free or reduced lunch status, or other variables used in the models. Our final analysis samples

---

[4]Students who scored zero on the open-ended mathematics test were outliers and influential in our results. We did not want our conclusions to be sensitive to a handful of unrepresentative students, so we excluded these students from our analysis.

included data on 1926 students with hands-on test scores and 1942 students with multiple-choice test scores.

Site 6

Our site 6 sample included all 25 middle schools in the district. Twelve schools had been involved in the USI for one or more years, and thirteen had not yet been involved. All 58 7th grade mathematics and 57 7th grade science teachers identified by the district completed surveys. However, only 57 of the responding mathematics teachers linked to seventh grade students and only 52 of the science teachers linked to students. The remaining respondents appear to be incorrectly identified as grade 7 teachers. The science sample included 3812 students, and 3279 of these had scores on the multiple-choice science test. The mathematics sample included 3682 students, of whom 3237 had scores on the multiple-choice mathematics test.

APPENDIX B

Items on Teaching Practices Scales

The wording of items is identical across sites.

The reform practices score for mathematics is the sum of scores on the 22 items listed in Table B1.

Table B1.

Items on Reform Practices Scale for Mathematics

About how often do you typically do each of the following in your *mathematics* instruction in this class?
Arrange seating to facilitate student discussion
Use open-ended questions
Require students to explain their reasoning when giving an answer
Encourage students to communicate mathematically
Encourage students to explore alternative methods for solutions
Allow students to work at their own pace
Read and comment on the reflections students have written in their notebooks or journals
About how often do students in this class typically take part in each of the following activities as part of their *mathematics* instruction?
Participate in student-led discussions
Work in cooperative learning groups
Make formal presentations to the class
Work on solving a real-world problem
Share ideas or solve problems with each other in small groups
Engage in hands-on mathematical activities
Design or implement their *own* investigations
Work on extended mathematics investigations (a week or more in duration)
Participate in field work
Record, represent, and/or analyze data
Write a description of a plan, procedure, or problem-solving process
Write reflections in a notebook or journal
Work on portfolios
Take tests requiring open-ended responses (e.g., descriptions, justifications of solutions)
Engage in performance tasks for assessment purposes

The traditional practices score for mathematics is the sum of scores on the 5 items listed in Table B2.

Table B2.

Items on Traditional Practices Scale for Mathematics

About how often do you typically do each of the following in your *mathematics* instruction in this class?
Lecture/introduce content through formal presentations
About how often do students in this class typically take part in each of the following activities as part of their *mathematics* instruction?
Read from a mathematics textbook in class
Practice computational skills
Memorize mathematics facts, rules, or formulas
Take short-answer tests (e.g., multiple-choice, true/false, fill-in-the-blank)

The reform practices score for science is the sum of scores on the 22 items listed in TableB3.

Table B3.

Items on Reform Practices Scale for Science

About how often do you typically do each of the following in your *science* instruction in this class?

- Arrange seating to facilitate student discussion
- Use open-ended questions
- Require students to supply evidence to support their claims
- Encourage students to explain concepts to one another
- Encourage students to consider alternative explanations
- Allow students to work at their own pace
- Read and comment on the reflections students have written in their notebooks or journals

About how often do students in this class typically take part in each of the following activities as part of their *science* instruction?

- Participate in student-led discussions
- Work in cooperative learning groups
- Make formal presentations to the class
- Work on solving a real-world problem
- Share ideas or solve problems with each other in small groups
- Engage in hands-on science activities
- Design or implement their *own* investigations
- Design objects within constraints (e.g., egg drop, toothpick bridge, aluminum boats)
- Work on extended science investigations or projects (a week or more in duration)
- Participate in field work
- Record, represent, and/or analyze data
- Write reflections in a notebook or journal
- Work on portfolios
- Take tests requiring open-ended responses (e.g., descriptions, justifications of solutions)
- Engage in performance tasks for assessment purposes

The traditional practices score for science is the sum of scores on the 5 items listed in TableB4.

Table B4.

Items on Traditional Practices Scale for Science

About how often do you typically do each of the following in your *science* instruction in this class?
Lecture/introduce content through formal presentations
About how often do students in this class typically take part in each of the following activities as part of their *science* instruction?
Read from a science textbook in class
Answer textbook/worksheet questions
Learn science vocabulary
Take short-answer tests (e.g., multiple-choice, true/false, fill-in-the-blank)

APPENDIX C

Full Regression Models

FULL REGRESSION MODELS FOR SITE 1

Table C1.

Regression Models for State Open-Ended Mathematics Test (Grade 4), Site 1

Model Including Reform Practices Scale Predictor Variable	Coefficient	T-statistic[c]	p-value
Intercept	0.289	3.130	0.002
Free or Reduced Lunches	-0.189	-2.856	0.004
African-American[a]	-0.140	-1.770	0.077
Hispanic	-0.252	-2.345	0.019
Other Race	-0.144	-1.267	0.205
Female	-0.023	-0.339	0.735
Reading Score	0.487	12.927	0.000
Reform Scale[b]	-0.010	-0.261	0.794

Model Including Traditional Practices Scale Predictor Variable	Coefficient	T-statistic[c]	p-value
Intercept	0.294	3.266	0.001
Free or Reduced Lunches	-0.192	-2.957	0.003
African-American[a]	-0.141	-1.806	0.071
Hispanic	-0.252	-2.378	0.017
Other Race	-0.144	-1.268	0.205
Female	-0.022	-0.332	0.740
Reading Score	0.486	12.832	0.000
Traditional Scale[b]	-0.006	-0.156	0.876

[a]Reference group for ethnicity is white, non-Hispanic
[b]Standardized to mean zero, standard deviation one
[c]T-statistics are adjusted for clustering (McCaffrey and Bell, 1997)

Table C2.

Regression Models for State Grid-In Mathematics Test (Grade 4), Site 1

Model Including Reform Practices Scale Predictor Variable	Coefficient	T-statistic[c]	p-value
Intercept	0.328	4.863	0.000
Free or Reduced Lunches	-0.159	-1.914	0.056
African-American[a]	-0.253	-3.730	0.000
Hispanic	-0.152	-1.423	0.155
Other Race	-0.090	-0.834	0.404
Female	-0.071	-1.098	0.272
Reading Score	0.516	16.964	0.000
Reform Scale[b]	0.050	1.941	0.052

Model Including Traditional Practices Scale Predictor Variable	Coefficient	T-statistic[c]	p-value
Intercept	0.324	4.749	0.000
Free or Reduced Lunches	-0.157	-1.872	0.061
African-American[a]	-0.252	-3.726	0.000
Hispanic	-0.166	-1.521	0.128
Other Race	-0.093	-0.870	0.385
Female	-0.072	-1.113	0.266
Reading Score	0.514	16.832	0.000
Traditional Scale[b]	-0.040	-1.792	0.073

[a]Reference group for ethnicity is white, non-Hispanic

[b]Standardized to mean zero, standard deviation one

[c]T-statistics are adjusted for clustering (McCaffrey and Bell, 1997)

Table C3.

Regression Models for State Multiple-Choice Mathematics Test (Grade 4), Site 1

Model Including Reform Practices Scale Predictor Variable	Coefficient	T-statistic[c]	p-value
Intercept	0.462	8.821	0.000
Free or Reduced Lunches	-0.206	-3.387	0.001
African-American[a]	-0.264	-3.709	0.000
Hispanic	-0.160	-1.986	0.047
Other Race	-0.218	-1.852	0.064
Female	-0.224	-4.099	0.000
Reading Score	0.631	19.294	0.000
Reform Scale[b]	0.031	0.870	0.385

Model Including Traditional Practices Scale Predictor Variable	Coefficient	T-statistic[c]	p-value
Intercept	0.463	9.206	0.000
Free or Reduced Lunches	-0.207	-3.303	0.001
African-American[a]	-0.265	-3.864	0.000
Hispanic	-0.172	-2.053	0.040
Other Race	-0.221	-1.872	0.061
Female	-0.224	-4.129	0.000
Reading Score	0.628	19.014	0.000
Traditional Scale[b]	-0.037	-1.223	0.221

[a]Reference group for ethnicity is white, non-Hispanic

[b]Standardized to mean zero, standard deviation one

[c]T-statistics are adjusted for clustering (McCaffrey and Bell, 1997)

FULL REGRESSION MODELS FOR SITE 2

Table C4.

Regression Models for RAND Hands-On Science Tasks (Grade 5), Site 2

Model Including Reform Practices Scale Predictor Variable	Coefficient	T-statistic[d]	p-value
Intercept	0.084	1.054	0.292
Free or Reduced Lunches	-0.169	-3.160	0.002
LEP	0.025	0.362	0.718
Gifted	0.369	5.790	0.000
Minority[a]	-0.105	-1.392	0.164
Female	0.121	2.568	0.010
Grade 4 State MC Mathematics[b]	0.307	9.360	0.000
Grade 4 State MC Reading[b]	0.306	10.014	0.000
Reform Scale[c]	0.113	4.066	0.000

Model Including Traditional Practices Scale Predictor Variable	Coefficient	T-statistic[d]	p-value
Intercept	0.105	1.337	0.181
Free or Reduced Lunches	-0.176	-3.130	0.002
LEP	0.007	0.088	0.930
Gifted	0.377	5.714	0.000
Minority[a]	-0.129	-1.658	0.097
Female	0.122	2.560	0.010
Grade 4 State MC Mathematics[b]	0.304	9.247	0.000
Grade 4 State MC Reading[b]	0.304	10.050	0.000
Traditional Scale[c]	0.030	0.946	0.344

[a]Includes all students not identified as white, non-Hispanic
[b]NCE standardized to mean zero, standard deviation one
[c]Standardized to mean zero, standard deviation one
[d]T-statistics are adjusted for clustering (McCaffrey and Bell, 1997)

Table C5.

Regression Models for SAT-9 Multiple-Choice Science Test (Grade 5), Site 2

Model Including Reform Practices Scale Predictor Variable	Coefficient	T-statistic[d]	p-value
Intercept	0.309	4.288	0.000
Free or Reduced Lunches	-0.110	-2.135	0.033
LEP	-0.070	-1.089	0.276
Gifted	0.351	5.949	0.000
Minority[a]	-0.191	-2.568	0.010
Female	-0.210	-4.351	0.000
Grade 4 State MC Mathematics[b]	0.159	4.414	0.000
Grade 4 State MC Reading[b]	0.349	10.841	0.000
Reform Scale[c]	0.053	2.036	0.042

Model Including Traditional Practices Scale Predictor Variable	Coefficient	T-statistic[d]	p-value
Intercept	0.309	3.999	0.000
Free or Reduced Lunches	-0.106	-1.999	0.046
LEP	-0.088	-1.369	0.171
Gifted	0.347	5.906	0.000
Minority[a]	-0.196	-2.469	0.014
Female	-0.208	-4.209	0.000
Grade 4 State MC Mathematics[b]	0.158	4.451	0.000
Grade 4 State MC Reading[b]	0.349	10.977	0.000
Traditional Scale[c]	0.049	1.865	0.062

[a]Includes all students not identified as white, non-Hispanic
[b]NCE standardized to mean zero, standard deviation one
[c]Standardized to mean zero, standard deviation one
[d]T-statistics are adjusted for clustering (McCaffrey and Bell, 1997)

Table C6.

Regression Models for SAT-9 Open-Ended Mathematics Test (Grade 5), Site 2

Model Including Reform Practices Scale Predictor Variable	Coefficient	T-statistic[d]	p-value
Intercept	0.117	1.043	0.297
Free or Reduced Lunches	-0.073	-1.293	0.196
LEP	-0.034	-0.402	0.688
Gifted	0.435	6.730	0.000
Minority[a]	-0.089	-0.885	0.376
Female	-0.092	-1.931	0.054
Grade 4 State MC Mathematics[b]	0.305	6.771	0.000
Grade 4 State MC Reading[b]	0.156	3.934	0.000
Reform Scale[c]	0.060	1.464	0.143

Model Including Traditional Practices Scale Predictor Variable	Coefficient	T-statistic[d]	p-value
Intercept	0.116	1.060	0.289
Free or Reduced Lunches	-0.075	-1.380	0.167
LEP	-0.029	-0.356	0.722
Gifted	0.436	6.924	0.000
Minority[a]	-0.089	-0.879	0.379
Female	-0.092	-1.934	0.053
Grade 4 State MC Mathematics[b]	0.308	6.785	0.000
Grade 4 State MC Reading[b]	0.148	3.706	0.000
Traditional Scale[c]	-0.012	-0.289	0.773

[a]Includes all students not identified as white, non-Hispanic
[b]NCE standardized to mean zero, standard deviation one
[c]Standardized to mean zero, standard deviation one
[d]T-statistics are adjusted for clustering (McCaffrey and Bell, 1997)

Table C7.

Regression State Multiple-Choice Mathematics Test (Grade 5), Site 2

Model Including Reform Practices Scale Predictor Variable	Coefficient	T-statistic[d]	p-value
Intercept	-0.075	-0.785	0.432
Free or Reduced Lunches	0.031	0.488	0.626
LEP	0.098	1.158	0.247
Gifted	0.216	3.789	0.000
Minority[a]	-0.026	-0.396	0.692
Female	0.024	0.562	0.574
Grade 4 State MC Mathematics[b]	0.563	12.817	0.000
Grade 4 State MC Reading[b]	0.132	3.877	0.000
Reform Scale[c]	0.020	0.471	0.637

Model Including Traditional Practices Scale Predictor Variable	Coefficient	T-statistic[d]	p-value
Intercept	-0.064	-0.699	0.484
Free or Reduced Lunches	0.020	0.312	0.755
LEP	0.066	0.801	0.423
Gifted	0.202	3.518	0.000
Minority[a]	-0.017	-0.269	0.788
Female	0.025	0.591	0.554
Grade 4 State MC Mathematics[b]	0.559	13.056	0.000
Grade 4 State MC Reading[b]	0.135	3.922	0.000
Traditional Scale[c]	0.094	2.359	0.018

[a]Includes all students not identified as white, non-Hispanic
[b]NCE standardized to mean zero, standard deviation one
[c]Standardized to mean zero, standard deviation one
[d]T-statistics are adjusted for clustering (McCaffrey and Bell, 1997)

FULL REGRESSION MODELS FOR SITE 3

Table C8.

Regression Models for CSIAC Hands-On Science Test (Grade 5), Site 3

Model Including Reform Practices Scale Predictor Variable	Coefficient	T-statistic[c]	p-value
Intercept	0.139	1.797	0.072
Free or Reduced Lunches	-0.106	-1.305	0.192
English Proficiency 1	0.062	0.289	0.774
English Proficiency 2	0.041	0.492	0.623
Special Education	-0.146	-1.435	0.151
African-American[a]	-0.074	-0.865	0.387
Hispanic	-0.155	-2.818	0.005
Asian	0.028	0.237	0.813
Female	0.205	4.376	0.000
Grade 4 Mathematics	0.172	4.232	0.000
Grade 4 Reading	0.352	7.731	0.000
Missing Grade 4 Scores	-0.160	-2.104	0.035
Reform Scale[b]	0.079	1.281	0.200
Model Including Traditional Practices Scale Predictor Variable	Coefficient	T-statistic[c]	p-value
Intercept	0.140	1.737	0.082
Free or Reduced Lunches	-0.113	-1.319	0.187
English Proficiency 1	0.069	0.320	0.751
English Proficiency 2	0.043	0.519	0.604
Special Education	-0.136	-1.350	0.177
African-American[a]	-0.065	-0.736	0.462
Hispanic	-0.145	-2.650	0.008
Asian	0.028	0.242	0.808
Female	0.209	4.394	0.000
Grade 4 Mathematics	0.175	4.273	0.000
Grade 4 Reading	0.358	7.684	0.000
Missing Grade 4 Scores	-0.153	-2.014	0.044
Traditional Scale[b]	0.075	1.328	0.184

[a]Reference group for ethnicity is white, non-Hispanic
[b]Standardized to mean zero, standard deviation one
[c]T-statistics are adjusted for clustering (McCaffrey and Bell, 1997)

Table C9.

Regression Models for CSIAC Multiple-Choice Science Test (Grade 5), Site 3

Model Including Reform Practices Scale Predictor Variable	Coefficient	T-statistic[c]	p-value
Intercept	0.473	8.291	0.000
Free or Reduced Lunches	-0.190	-3.388	0.001
English Proficiency 1	-0.199	-0.855	0.405
English Proficiency 2	-0.057	-0.804	0.421
Special Education	0.022	0.336	0.737
African-American[a]	-0.248	-3.336	0.001
Hispanic	-0.193	-3.549	0.000
Asian	-0.239	-2.756	0.006
Female	-0.078	-1.714	0.087
Grade 4 Mathematics	0.135	3.813	0.000
Grade 4 Reading	0.521	16.222	0.000
Missing Grade 4 Scores	-0.128	-2.389	0.017
Reform Scale[b]	0.049	1.951	0.051
Model Including Traditional Practices Scale Predictor Variable	Coefficient	T-statistic[c]	p-value
Intercept	0.480	8.247	0.000
Free or Reduced Lunches	-0.202	-3.568	0.000
English Proficiency 1	-0.192	-0.821	0.423
English Proficiency 2	-0.057	-0.806	0.420
Special Education	0.024	0.377	0.706
African-American[a]	-0.242	-3.286	0.001
Hispanic	-0.193	-3.470	0.001
Asian	-0.241	-2.746	0.006
Female	-0.078	-1.709	0.088
Grade 4 Mathematics	0.136	3.806	0.000
Grade 4 Reading	0.526	16.461	0.000
Missing Grade 4 Scores	-0.124	-2.321	0.021
Traditional Scale[b]	0.007	0.280	0.779

[a]Reference group for ethnicity is white, non-Hispanic
[b]Standardized to mean zero, standard deviation one
[c]T-statistics are adjusted for clustering (McCaffrey and Bell, 1997)

Table C10.

Regression Models for SAT-9 Open-Ended Mathematics Test (Grade 5), Site 3

Model Including Reform Practices Scale Predictor Variable	Coefficient	T-statistic[c]	p-value
Intercept	0.356	3.454	0.001
Free or Reduced Lunches	-0.106	-1.171	0.242
English Proficiency 1	0.005	0.037	0.970
English Proficiency 2	0.134	2.121	0.034
Special Education	-0.204	-2.510	0.012
African-American[a]	-0.331	-4.526	0.000
Hispanic	-0.183	-2.915	0.004
Asian	0.026	0.262	0.794
Female	-0.168	-3.407	0.001
Grade 4 Mathematics	0.277	5.531	0.000
Grade 4 Reading	0.275	6.424	0.000
Missing Grade 4 Scores	-0.137	-2.296	0.022
Reform Scale[b]	0.092	2.055	0.040

Model Including Traditional Practices Scale Predictor Variable	Coefficient	T-statistic[c]	p-value
Intercept	0.374	3.606	0.000
Free or Reduced Lunches	-0.125	-1.386	0.166
English Proficiency 1	0.010	0.070	0.944
English Proficiency 2	0.135	2.104	0.035
Special Education	-0.219	-2.683	0.008
African-American[a]	-0.336	-4.328	0.000
Hispanic	-0.192	-2.914	0.004
Asian	0.043	0.412	0.680
Female	-0.175	-3.560	0.000
Grade 4 Mathematics	0.281	5.183	0.000
Grade 4 Reading	0.287	6.619	0.000
Missing Grade 4 Scores	-0.157	-2.489	0.013
Traditional Scale[b]	-0.091	-1.760	0.078

[a]Reference group for ethnicity is white, non-Hispanic
[b]Standardized to mean zero, standard deviation one
[c]T-statistics are adjusted for clustering (McCaffrey and Bell, 1997)

Table C11.

Regression Models for CTBS Multiple-Choice Mathematics Test (Grade 5) Total Score, Site 3

Model Including Reform Practices Scale Predictor Variable	Coefficient	T-statistic[c]	p-value
Intercept	0.360	4.491	0.000
Free or Reduced Lunches	-0.176	-2.436	0.015
English Proficiency 1	-0.036	-0.240	0.812
English Proficiency 2	0.102	1.850	0.064
Special Education	-0.228	-3.149	0.002
African-American[a]	-0.080	-1.323	0.186
Hispanic	-0.120	-2.318	0.021
Asian	0.044	0.532	0.595
Female	-0.089	-2.059	0.040
Grade 4 Mathematics	0.365	9.004	0.000
Grade 4 Reading	0.380	13.073	0.000
Missing Grade 4 Scores	-0.132	-2.377	0.018
Reform Scale[b]	0.044	1.047	0.295

Model Including Traditional Practices Scale Predictor Variable	Coefficient	T-statistic[c]	p-value
Intercept	0.366	4.798	0.000
Free or Reduced Lunches	-0.182	-2.464	0.014
English Proficiency 1	-0.037	-0.252	0.803
English Proficiency 2	0.102	1.868	0.062
Special Education	-0.233	-3.195	0.002
African-American[a]	-0.080	-1.288	0.198
Hispanic	-0.121	-2.245	0.025
Asian	0.055	0.640	0.522
Female	-0.093	-2.219	0.027
Grade 4 Mathematics	0.368	8.828	0.000
Grade 4 Reading	0.386	13.142	0.000
Missing Grade 4 Scores	-0.143	-2.487	0.014
Traditional Scale[b]	-0.050	-1.460	0.144

[a]Reference group for ethnicity is white, non-Hispanic
[b]Standardized to mean zero, standard deviation one
[c]T-statistics are adjusted for clustering (McCaffrey and Bell, 1997)

Table C12.

Regression Models for CTBS Multiple-Choice Mathematics Test (Grade 5) General Mathematics Score, Site 3

Model Including Reform Practices Scale Predictor Variable	Coefficient	T-statistic[c]	p-value
Intercept	0.423	5.408	0.000
Free or Reduced Lunches	-0.169	-2.549	0.011
English Proficiency 1	-0.096	-0.621	0.538
English Proficiency 2	0.052	0.899	0.369
Special Education	-0.197	-2.590	0.010
African-American[a]	-0.107	-1.967	0.049
Hispanic	-0.136	-2.501	0.013
Asian	-0.028	-0.352	0.725
Female	-0.113	-2.592	0.010
Grade 4 Mathematics	0.328	8.485	0.000
Grade 4 Reading	0.399	14.430	0.000
Missing Grade 4 Scores	-0.172	-3.193	0.002
Reform Scale[b]	0.048	1.234	0.217

Model Including Traditional Practices Scale Predictor Variable	Coefficient	T-statistic[c]	p-value
Intercept	0.429	5.837	0.000
Free or Reduced Lunches	-0.173	-2.553	0.011
English Proficiency 1	-0.101	-0.675	0.504
English Proficiency 2	0.050	0.862	0.389
Special Education	-0.205	-2.668	0.008
African-American[a]	-0.109	-1.919	0.055
Hispanic	-0.138	-2.429	0.015
Asian	-0.016	-0.205	0.838
Female	-0.118	-2.802	0.005
Grade 4 Mathematics	0.332	8.443	0.000
Grade 4 Reading	0.406	14.577	0.000
Missing Grade 4 Scores	-0.187	-3.362	0.001
Traditional Scale[b]	-0.070	-2.130	0.033

[a]Reference group for ethnicity is white, non-Hispanic
[b]Standardized to mean zero, standard deviation one
[c]T-statistics are adjusted for clustering (McCaffrey and Bell, 1997)

Table C13.

Regression Models for CTBS Multiple-Choice Mathematics Test (Grade 5) Computation Score, Site 3

Model Including Reform Practices Scale Predictor Variable	Coefficient	T-statistic[c]	p-value
Intercept	0.109	0.975	0.330
Free or Reduced Lunches	-0.153	-1.591	0.112
English Proficiency 1	0.128	0.843	0.403
English Proficiency 2	0.208	3.125	0.002
Special Education	-0.252	-3.238	0.001
African-American[a]	0.010	0.123	0.902
Hispanic	-0.049	-0.696	0.486
Asian	0.222	1.956	0.051
Female	-0.004	-0.086	0.931
Grade 4 Mathematics	0.375	7.743	0.000
Grade 4 Reading	0.240	5.980	0.000
Missing Grade 4 Scores	0.004	0.069	0.945
Reform Scale[b]	0.023	0.427	0.669

Model Including Traditional Practices Scale Predictor Variable	Coefficient	T-statistic[c]	p-value
Intercept	0.113	0.987	0.323
Free or Reduced Lunches	-0.162	-1.611	0.107
English Proficiency 1	0.139	0.901	0.371
English Proficiency 2	0.213	3.282	0.001
Special Education	0.373	7.528	0.000
African-American[a]	0.243	6.025	0.000
Hispanic	0.005	0.082	0.935
Asian	-0.251	-3.221	0.001
Female	0.013	0.157	0.875
Grade 4 Mathematics	-0.048	-0.694	0.488
Grade 4 Reading	0.227	1.975	0.048
Missing Grade 4 Scores	-0.005	-0.095	0.925
Traditional Scale[b]	0.014	0.255	0.799

[a]Reference group for ethnicity is white, non-Hispanic
[b]Standardized to mean zero, standard deviation one
[c]T-statistics are adjusted for clustering (McCaffrey and Bell, 1997)

FULL REGRESSION MODELS FOR SITE 4

Table C14.

Regression Models for CSIAC Hands-On Science Test (Grade 5), Site 4

Model Including Reform Practices Scale Predictor Variable	Coefficient	T-statistic[c]	p-value
Intercept	0.192	2.107	0.035
Free or Reduced Lunches	-0.086	-0.814	0.415
English Proficiency 1	-0.401	-2.662	0.008
English Proficiency 2	0.446	2.557	0.011
Gifted	0.207	3.101	0.002
Special Education	0.002	0.018	0.986
African-American[a]	-0.875	-6.689	0.000
Chinese	-0.032	-0.359	0.719
Filipino	-0.317	-1.809	0.071
Hispanic	-0.097	-0.776	0.438
Japanese	-0.079	-0.420	0.675
Korean	0.176	1.166	0.243
Other Race/Ethnicity	-0.183	-1.509	0.131
Female	0.066	1.023	0.306
Grade 4 Mathematics	0.138	1.643	0.100
Grade 4 Reading	0.193	2.142	0.032
Reform Scale[b]	0.047	0.707	0.480

[a]Reference group for ethnicity is white, non-Hispanic
[b]Standardized to mean zero, standard deviation one
[c]T-statistics are adjusted for clustering (McCaffrey and Bell, 1997)

Table C15.

Regression Models for CSIAC Hands-On Science Test (Grade 5), Site 4 (Continued)

Model Including Traditional Practices Scale			
Predictor Variable	Coefficient	T-statistic[c]	p-value
Intercept	0.191	2.060	0.039
Free or Reduced Lunches	-0.079	-0.756	0.450
English Proficiency 1	-0.401	-2.611	0.009
English Proficiency 2	0.460	2.655	0.008
Gifted	0.205	3.041	0.002
Special Education	-0.006	-0.044	0.965
African-American[a]	-0.860	-6.354	0.000
Chinese	-0.024	-0.269	0.788
Filipino	-0.299	-1.677	0.094
Hispanic	-0.098	-0.792	0.428
Japanese	-0.044	-0.234	0.815
Korean	0.188	1.236	0.216
Other Race/Ethnicity	-0.168	-1.337	0.181
Female	0.064	1.019	0.308
Grade 4 Mathematics	0.138	1.664	0.096
Grade 4 Reading	0.192	2.160	0.031
Traditional Scale[b]	-0.008	-0.137	0.891

[a]Reference group for ethnicity is white, non-Hispanic
[b]Standardized to mean zero, standard deviation one
[c]T-statistics are adjusted for clustering (McCaffrey and Bell, 1997)

Table C16.

Regression Models for CSIAC Multiple-Choice Science Test (Grade 5), Site 4

Model Including Reform Practices Scale			
Predictor Variable	Coefficient	T-statistic[c]	p-value
Intercept	0.209	2.570	0.010
Free or Reduced Lunches	-0.144	-2.384	0.017
English Proficiency 1	-0.480	-5.276	0.000
English Proficiency 2	0.221	1.229	0.219
Gifted	0.334	5.477	0.000
Special Education	0.230	1.933	0.053
African-American[a]	-0.676	-7.262	0.000
Chinese	0.004	0.058	0.953
Filipino	-0.240	-2.011	0.045
Hispanic	-0.029	-0.280	0.780
Japanese	0.027	0.112	0.911
Korean	0.352	3.262	0.001
Other Race/Ethnicity	-0.115	-1.607	0.108
Female	-0.218	-4.823	0.000
Grade 4 Mathematics	0.112	1.977	0.048
Grade 4 Reading	0.415	7.987	0.000
Reform Scale[b]	-0.002	-0.056	0.955

[a]Reference group for ethnicity is white, non-Hispanic
[b]Standardized to mean zero, standard deviation one
[c]T-statistics are adjusted for clustering (McCaffrey and Bell, 1997)

Table C17.

Regression Models for CSIAC Multiple-Choice Science Test (Grade 5), Site 4 (Continued)

Model Including Traditional Practices Scale			
Predictor Variable	Coefficient	T-statistic[c]	p-value
Intercept	0.132	1.675	0.094
Free or Reduced Lunches	-0.145	-2.763	0.006
English Proficiency 1	-0.475	-5.217	0.000
English Proficiency 2	0.248	1.397	0.162
Gifted	0.325	5.400	0.000
Special Education	0.232	1.984	0.047
African-American[a]	-0.695	-7.202	0.000
Chinese	0.009	0.150	0.881
Filipino	-0.240	-2.070	0.039
Hispanic	-0.048	-0.464	0.643
Japanese	0.067	0.273	0.785
Korean	0.349	3.247	0.001
Other Race/Ethnicity	-0.112	-1.609	0.108
Female	-0.204	-4.400	0.000
Grade 4 Mathematics	0.119	2.250	0.024
Grade 4 Reading	0.404	8.290	0.000
Traditional Scale[b]: Linear	-0.056	-1.266	0.206
Traditional Scale[b]: Quadratic	0.078	2.017	0.044

[a]Reference group for ethnicity is white, non-Hispanic
[b]Standardized to mean zero, standard deviation one
[c]T-statistics are adjusted for clustering (McCaffrey and Bell, 1997)

FULL REGRESSION MODELS FOR SITE 5

Table C18.

Regression Models for RAND Hands-On Science Tasks (Grade 7), Site 5

Model Including Reform Practices Scale Predictor Variable	Coefficient	T-statistic[c]	p-value
Intercept	3.225	6.307	0.000
Free or Reduced Lunches	-0.276	-4.706	0.000
Age	-0.231	-5.813	0.000
Minority[a]	-0.753	-5.947	0.000
Female	0.082	1.743	0.081
Reform Scale[b]	0.033	0.364	0.716

Model Including Traditional Practices Scale Predictor Variable	Coefficient	T-statistic[c]	p-value
Intercept	3.162	6.629	0.000
Free or Reduced Lunches	-0.220	-4.079	0.000
Age	-0.232	-6.321	0.000
Minority[a]	-0.739	-7.324	0.000
Female	0.072	1.665	0.096
Traditional Scale[b]	-0.171	-3.762	0.000

[a]Includes all students not identified as white, non-Hispanic
[b]Standardized to mean zero, standard deviation one
[c]T-statistics are adjusted for clustering (McCaffrey and Bell, 1997)

Table C19.

Regression Models for SAT-9 Multiple-Choice Science Tests (Grade 7), Site 5

Model Including Reform Practices Scale Predictor Variable	Coefficient	T-statistic[c]	p-value
Intercept	3.351	5.280	0.000
Free or Reduced Lunches	-0.269	-4.440	0.000
Age	-0.234	-4.801	0.000
Minority[a]	-0.790	-4.529	0.000
Female	-0.086	-1.278	0.201
Reform Scale[b]	0.083	0.782	0.434

Model Including Traditional Practices Scale Predictor Variable	Coefficient	T-statistic[c]	p-value
Intercept	3.263	4.671	0.000
Free or Reduced Lunches	-0.254	-4.433	0.000
Age	-0.229	-4.276	0.000
Minority[a]	-0.791	-4.370	0.000
Female	-0.086	-1.238	0.216
Traditional Scale[b]	-0.084	-1.397	0.163

[a]Includes all students not identified as white, non-Hispanic
[b]Standardized to mean zero, standard deviation one
[c]T-statistics are adjusted for clustering (McCaffrey and Bell, 1997)

Table C20.

Regression Models for SAT-9 Open-Ended Mathematics Tests (Grade 7), Site 5

Model Including Reform Practices Scale Predictor Variable	Coefficient	T-statistic[c]	p-value
Intercept	3.539	9.815	0.000
Free or Reduced Lunches	-0.129	-3.027	0.002
Age	-0.256	-9.445	0.000
Minority[a]	-0.505	-11.333	0.000
Female	-0.062	-1.809	0.070
Reform Scale[b]	0.108	2.727	0.006

Model Including Traditional Practices Scale Predictor Variable	Coefficient	T-statistic[c]	p-value
Intercept	3.657	9.779	0.000
Free or Reduced Lunches	-0.139	-3.363	0.001
Age	-0.265	-9.196	0.000
Minority[a]	-0.517	-11.628	0.000
Female	-0.064	-1.833	0.067
Traditional Scale[b]	-0.056	-1.572	0.116

[a]Includes all students not identified as white, non-Hispanic
[b]Standardized to mean zero, standard deviation one
[c]T-statistics are adjusted for clustering (McCaffrey and Bell, 1997)

Table C21.

Regression Models for State Multiple-Choice Mathematics Tests (Grade 7), Site 5

Model Including Reform Practices Scale Predictor Variable	Coefficient	T-statistic[c]	p-value
Intercept	4.257	7.673	0.000
Free or Reduced Lunches	-0.169	-3.027	0.002
Age	-0.316	-8.173	0.000
Minority[a]	-0.386	-6.710	0.000
Female	-0.054	-1.618	0.106
Reform Scale[b]	-0.004	-0.055	0.956

Model Including Traditional Practices Scale Predictor Variable	Coefficient	T-statistic[c]	p-value
Intercept	4.240	8.150	0.000
Free or Reduced Lunches	-0.165	-3.180	0.001
Age	-0.315	-8.683	0.000
Minority[a]	-0.383	-6.544	0.000
Female	-0.054	-1.631	0.103
Traditional Scale[b]	-0.018	-0.244	0.807

[a]Includes all students not identified as white, non-Hispanic
[b]Standardized to mean zero, standard deviation one
[c]T-statistics are adjusted for clustering (McCaffrey and Bell, 1997)

FULL REGRESSION MODELS FOR SITE 6

Table C22.

Regression Models for MAT-7 Multiple-Choice Science Test (Grade 7), Site 6

Model Including Reform Practices Scale Predictor Variable	Coefficient	T-statistic[d]	p-value
Intercept	0.146	3.472	0.001
Free or Reduced Lunches	-0.188	-5.729	0.000
Minority[a]	-0.062	-1.948	0.051
Female	-0.053	-1.758	0.079
Grade 6 Science[b]	0.620	26.376	0.000
Reform Scale[c]	0.051	2.750	0.006

Model Including Traditional Practices Scale Predictor Variable	Coefficient	T-statistic[d]	p-value
Intercept	0.149	3.351	0.001
Free or Reduced Lunches	-0.198	-6.071	0.000
Minority[a]	-0.051	-1.620	0.105
Female	-0.053	-1.775	0.076
Grade 6 Science[b]	0.624	26.640	0.000
Traditional Scale[c]	-0.004	-0.149	0.881

[a]Includes all students not identified as white, non-Hispanic
[b]Scale score standardized to mean zero, standard deviation one
[c]Standardized to mean zero, standard deviation one
[d]T-statistics are adjusted for clustering (McCaffrey and Bell, 1997)

Table C23.

Regression Models for MAT-7 Multiple-Choice Mathematics Test Total Scores (Grade 7), Site 6

Model Including Reform Practices Scale Predictor Variable	Coefficient	T-statistic[d]	p-value
Intercept	0.097	1.867	0.062
Free or Reduced Lunches	-0.076	-1.917	0.055
Minority[a]	-0.048	-1.185	0.236
Female	-0.031	-1.308	0.191
Grade 6 Mathematics[b]	0.755	24.564	0.000
Missing Grade 6 Mathematics	-0.148	-3.598	0.000
Reform Scale[c]	0.046	1.589	0.112

Model Including Traditional Practices Scale Predictor Variable	Coefficient	T-statistic[d]	p-value
Intercept	0.105	1.963	0.050
Free or Reduced Lunches	-0.071	-1.695	0.090
Minority[a]	-0.058	-1.389	0.165
Female	-0.033	-1.393	0.164
Grade 6 Mathematics[b]	0.752	23.682	0.000
Missing Grade 6 Mathematics	-0.154	-3.724	0.000
Traditional Scale[c]	-0.004	-0.096	0.923

[a]Includes all students not identified as white, non-Hispanic
[b]Scale score standardized to mean zero, standard deviation one
[c]Standardized to mean zero, standard deviation one
[d]T-statistics are adjusted for clustering (McCaffrey and Bell, 1997)

Table C24.

Regression Models for MAT-7 Mathematics Concepts Subtest (Grade 7), Site 6

Model Including Reform Practices Scale Predictor Variable	Coefficient	T-statistic[d]	p-value
Intercept	0.133	3.161	0.002
Free or Reduced Lunches	-0.061	-1.717	0.086
Minority[a]	-0.081	-2.169	0.030
Female	-0.075	-3.069	0.002
Grade 6 Mathematics[b]	0.734	28.298	0.000
Missing Grade 6 Mathematics	-0.135	-3.408	0.001
Reform Scale[c]	0.062	2.650	0.008

Model Including Traditional Practices Scale Predictor Variable	Coefficient	T-statistic[d]	p-value
Intercept	0.144	3.239	0.001
Free or Reduced Lunches	-0.055	-1.412	0.158
Minority[a]	-0.095	-2.396	0.017
Female	-0.078	-3.218	0.001
Grade 6 Mathematics[b]	0.729	27.398	0.000
Missing Grade 6 Mathematics	-0.143	-3.628	0.000
Traditional Scale[c]	-0.007	-0.214	0.830

[a]Includes all students not identified as white, non-Hispanic
[b]Scale score standardized to mean zero, standard deviation one
[c]Standardized to mean zero, standard deviation one
[d]T-statistics are adjusted for clustering (McCaffrey and Bell, 1997)

Table C25.

Regression Models for MAT-7 Mathematics Procedures Subtest (Grade 7), Site 6

Model Including Reform Practices Scale Predictor Variable	Coefficient	T-statistic[d]	p-value
Intercept	0.043	0.688	0.492
Free or Reduced Lunches	-0.088	-1.964	0.050
Minority[a]	0.002	0.038	0.970
Female	0.037	1.360	0.174
Grade 6 Mathematics[b]	-0.151	-3.043	0.003
Missing Grade 6 Mathematics	0.673	20.440	0.000
Reform Scale[c]	0.017	0.429	0.668

Model Including Traditional Practices Scale Predictor Variable	Coefficient	T-statistic[d]	p-value
Intercept	0.047	0.733	0.463
Free or Reduced Lunches	-0.086	-1.872	0.061
Minority[a]	-0.002	-0.045	0.964
Female	0.036	1.316	0.188
Grade 6 Mathematics[b]	0.672	20.045	0.000
Missing Grade 6 Mathematics	-0.154	-3.042	0.003
Traditional Scale[c]	-0.003	-0.073	0.942

[a]Includes all students not identified as white, non-Hispanic
[b]Scale score standardized to mean zero, standard deviation one
[c]Standardized to mean zero, standard deviation one
[d]T-statistics are adjusted for clustering (McCaffrey and Bell, 1997)

APPENDIX D

Details on Pooled Analysis of Regression Coefficients

We used the following model to obtain our pooled estimates and to determine if sites were comparable and pooled analyses appropriate. For a given subject, test type and scale, we assume that our sample denotes a random sample of all possible sites and that

$$b_i = \beta + \eta_i + \varepsilon_i ,$$

where b_i denotes the estimated coefficient from the ith site, β denotes the average coefficient across all sites, η_i denotes the deviation of site i from the average and ε_i denotes sampling error in our estimate b_i as an estimate of $\beta + \eta_i$. The deviations, η_i are assumed to be normally distributed with mean zero and variance τ^2. The errors, ε_i are assumed to be normally distributed with mean zero and variance σ^2_i. Error variability differs from site to site depending on the distribution of teacher responses and other covariates and depending on the residual variance from the regression model fit in each site.

We used the square of the standard error estimates from the individual site analyses as our estimates of the σ^2_i parameters. Treating these estimates as fixed we then estimate τ^2 using restricted maximum likelihood (Searle , Cassella and McCullogh, 199?). This method also provides a confidence interval for our estimate of τ^2. We then estimated the average coefficient, β, as a weighted average of the b_i's, where the weight for the ith site is

$$w_i = \frac{1}{\hat{\tau}^2 + \hat{\sigma}_i^2} \Big/ \sum_j \frac{1}{\hat{\tau}^2 + \hat{\sigma}_j^2}.$$

The weighted averages, their standard errors, and lower and upper 95 percent confidence bounds are provided in Table D1. Estimates and confidence intervals of τ^2 are also given for each set of pooled analyses.

Table D1.

Results from Pooled Analyses of Relationships Between Practices and Achievement

Subject	Test Format	Scale	Weighted Average	Standard Error	CILB	CIUB	Tau^2	CI
Mathematics	MC	Reform	0.035	0.018	0.001	0.070	0.000	(0, 0.003)
Mathematics	OR	Reform	0.060	0.028	0.006	0.115	0.001	(0, 0.024)
Mathematics	MC	Traditional	-0.004	0.027	-0.057	0.049	0.002	(0, 0.020)
Mathematics	OR	Traditional	-0.037	0.020	-0.077	0.003	0.000	(0, 0.011)
Science	MC	Reform	0.046	0.012	0.022	0.070	0.000	(0, 0.003)
Science	OR	Reform	0.095	0.023	0.050	0.140	0.000	(0, 0.016)
Science	MC	Traditional	0.009	0.017	-0.023	0.042	0.000	(0, 0.015)
Science	OR	Traditional	-0.020	0.054	-0.126	0.086	0.009	(.001,0.099)

APPENDIX E

Results from Analysis of Format Differences

Table E1.

Standardized Coefficients for Models Predicting Differences Between Formats

Site	Subject	Scale	Beta	Standard Error
1	Mathematics	Reform	-0.041	0.028
1	Mathematics	Traditional	0.032	0.034
2	Mathematics	Reform	0.048	0.049
2	Mathematics	Traditional	-0.111	0.051
2	Science	Reform	0.049	0.032
2	Science	Traditional	-0.017	0.035
3	Mathematics	Reform	0.051	0.040
3	Mathematics	Traditional	-0.028	0.046
3	Science	Reform	0.045	0.070
3	Science	Traditional	0.071	0.064
4	Science	Reform	0.046	0.058
4	Science	Traditional	0.000	0.047
5	Mathematics	Reform	0.119	0.058
5	Mathematics	Traditional	-0.040	0.067
5	Science	Reform	-0.048	0.068
5	Science	Traditional	-0.102	0.045

Table E2.

Results from Pooled Analyses of Differences Between Formats

Subject	Scale	Weighted Average	Standard Error	CILB	CIUB	Tau^2	CI
Mathematics	Reform	0.033	0.034	-0.034	0.101	0.003	(0,0.0377)
Mathematics	Traditional	-0.029	0.033	-0.093	0.036	0.002	(0,0.0325)
Science	Reform	0.035	0.024	-0.012	0.048	0.000	(0,0.0185)
Science	Traditional	-0.020	0.030	-0.078	0.039	0.002	(0,0.0350)

REFERENCES

American Association for the Advancement of Science (1993). *Benchmarks for science literacy: Project 2061.* New York: Oxford University Press.

Bond, L. A., Braskamp, D., and van der Ploeg, A. (1996). *State student assessment programs data base.* Oak Brook, IL: North Central Regional Educational Laboratory.

Cohen, D. K., & Ball, D. L. (1990). Relations between policy and practice: A commentary. *Educational Evaluation and Policy Analysis, 12,* 331-338.

Cohen, D. K., & Hill, H. C. (1998). *State policy and classroom performance: Mathematics reform in California* (CPRE Policy Brief). Philadelphia: Consortium for Policy Research in Education.

Consortium for Policy Research in Education (1995a). *Reforming science, mathematics, and technology education: NSF's State Systemic Initiatives* (CPRE Policy Brief). New Brunswick, NJ: Author.

Consortium for Policy Research in Education (1995b). *Tracking student achievement in science and math: The promise of state assessment programs* (CPRE Policy Brief). New Brunswick, NJ: Author.

Corcoran, T. B., Shields, P. M., and Zucker, A. A. (1998). *Evaluation of the National Science Foundation's Statewide Systemic Initiatives (SSI) Program: The SSI's and Professional Development for Teachers.* Menlo Park: SRI International.

Fantuzzo, J. W., King, J. A., & Heller, L. R. (1992). Effects of reciprocal peer tutoring on mathematics and school adjustment: A component analysis. *Journal of Educational Psychology, 84,* 331-339.

Fox, J. (1998). *NSF programs attacked as weak, unclear.* Education Daily, July 24, pp. 1-2.

Ginsburg-Block, M. D., & Fantuzzo, J. W. (1998). An evaluation of the relative effectiveness of NCTM standards-based interventions for low-achieving urban elementary students. *Journal of Educational Psychology, 90,* 560-569.

Goldstein, H. (1995). Multilevel Statistical Models Second Edition. London: Arnold.

Greenwood, C. R., Carta, J. J., & Hall, R. V. (1988). The use of peer tutoring strategies in classroom management and educational instruction. *School Psychology Review, 17,* 258-275.

Hill, P. T. (1995). *Reinventing public education.* Santa Monica, CA: RAND.

Knapp, M. S. (1997). Between systemic reforms and the mathematics and science classroom: The dynamics of innovation, implementation, and professional learning. *Review of Educational Research, 67,* 227-266.

Koretz, D., and Barron, S. I. (1998, in press). *The validity of gains on the Kentucky Instructional Results Information System (KIRIS).* Santa Monica: RAND.

Koretz, D., Linn, R.L., Dunbar, S.B., and Shepard, L.A. (1991). The Effects of High-Stakes Testing: Preliminary Evidence About Generalization Across Tests, in R.L. Linn (chair), *The Effects of High Stakes Testing*, symposium presented at the annual

meetings of the American Educational Research Association and the National Council on Measurement in Education, Chicago, April.

Laguarda, K. G. (1998). *Assessing the SSIs' impacts on student achievement: An imperfect science.* Menlo Park, CA: SRI International.

Mayer, D. P. (1998). Do new teaching standards undermine performance on old tests? *Educational Evaluation and Policy Analysis, 20,* 53-73.

McCaffrey, D., & Bell, R. (1997). *Bias reduction in standard error estimates for regression analyses from multi-stage designs with few primary sampling units.* Paper presented at the Joint Statistical Meetings, Anaheim CA.

National Council of Teachers of Mathematics (1989). *Curriculum and evaluation standards for school mathematics.* Reston, VA: Author.

National Research Council (1996). *National science education standards.* Washington, DC: National Academy Press.

Searle , Cassella and McCullogh, 199? (GET REFERENCE FROM DAN)

Shields, P. M., Corcoran, T. B., & Zucker, A. A. (1994). *Evaluation of NSF's Statewide Systemic Initiatives (SSI) program: First-year report.* Menlo Park, CA: SRI International.

Shields, P. M., Marsh, J. A., and Adelman, N. E. (1998). *Evaluation of the National Science Foundation's Statewide Systemic Initiatives (SSI) Program: The SSI's Impacts on Classroom Practice.* Menlo Park: SRI International.

Smith, M., & O'Day, J. (1991). Systemic school reform. In S. H. Fuhrman & B. Malen (Eds.), *The politics of curriculum and testing* (pp. 233-268). Bristol, PA: The Falmer Press.

Stecher, B. M., & Klein, S. P. (Eds.) (1996). *Performance assessments in science: Hands-on tasks and scoring guides* (MR-660-NSF). Santa Monica, CA: RAND.

Tyack, D., & Cuban, L. (1995). *Tinkering toward utopia*. Cambridge, MA: Harvard University Press.

Verschaffel, L., & De Corte, E. (1997). Teaching realistic mathematical modeling in the elementary school: A teaching experiment with fifth-graders. *Journal for Research in Mathematics Education, 28,* 577-601.

Webb, N. M., & Palincsar, A. S. (1996). Group processes in the classroom. In D. C. Berliner & R. C. Calfee (Eds.), *Handbook of educational psychology* (pp. 841-873). New York: Macmillan.

Weiss, I. R., Montgomery, D. L., Ridgway, C. J., & Bond, S. L. (1998). *Local Systemic Change through Teacher Enhancement: Year three cross-site report*. Chapel Hill, NC: Horizon Research, Inc.

Williams, L. (1998). *The Urban Systemic Initiatives (USI) program of the National Science Foundation: Summary update*. Washington, DC: NSF.